Efectos de la Imputación en el Análisis de Datos Multivariados

Marcia Gabriela Cuenca Calle[1] , Gaudencio Zurita Herrera[2]

[1]Ingeniera en Estadística e Informática 2006, e-mail: mcuenca@espol.edu.ec
[2]Director del Centro de Estudios e Investigaciones Estadísticas ICM – ESPOL, M.Sc. en Matemáticas, 1974, M.Sc. en Estadística, 1981, University of South Carolina. Profesor de ESPOL desde 1969., e-mail: gzurita@espol.edu.ec

Resumen

El presente trabajo consiste en un análisis estadístico de los Efectos de la Imputación en el Análisis de Datos Multivariados, basados en la generación de muestras con variables aleatorias dependientes e independientes de diferentes tamaños y distribuciones. El objetivo de este estudio es el comparar los Métodos de Imputación para el manejo de datos incompletos, tales como: Imputación por la Media e Imputación por Regresión, utilizando diferentes tamaños de muestras: 30, 50 y 100 y distribuciones tales como: Normal, Poisson y Exponencial, con el fin de comprobar que método de imputación brinda resultados de predicción que tiendan al dato observado.

Palabras Claves: *Imputación, Datos Multivariados*

Abstract

The present work consists of a statistical analysis of the Effects of the Imputation in the Analysis of Multivariate Data, based on the generation of samples with dependent and independent variates of different sizes and distributions. The objective of this study is to compare the Methods of Imputation for the handling of incompletos data, such as: Imputation by the Average and Imputation by Regression, using different sizes from samples: 30, 50 and 100 and distributions such as: Normal, Poisson and Exponential, with the purpose of verifying that imputation method offers prediction results that tend to the observed data.

1. Introducción

Este trabajo se basa en un análisis acerca de los Efectos de la Imputación en el Análisis de Datos Multivariados. Este estudio tiene como objetivo, comparar los Métodos de Imputación más usados en la actualidad, tales como: Imputación por la Media e Imputación por Regresión. Para lograr este objetivo se utiliza métodos estadísticos univariados y multivariados como: estadística descriptiva, diagrama de cajas, matrices de varianzas y covarianzas y matrices de correlaciones, entre otros que se consideraron necesarios para el desarrollo de este trabajo. En base a los resultados obtenidos se establecen las conclusiones y recomendaciones.

2. La Pérdida de Datos en una Investigación

En el análisis de datos reales es habitual encontrarse con matrices que tienen sus datos incompletos ya sea por inconvenientes en la recolección de la información, por la negativa a cooperar, incapacidad de contestar de los entrevistados, ausencia temporal del entrevistado, pérdida de formularios, errores de digitalización, etc.

Esta situación dificulta el tratamiento y análisis de los datos así como también la utilización de los procedimientos estadísticos estándares ya que estamos dentro de un problema de falta de datos, lo cual puede introducir sesgo en la estimación e incrementar o disminuir la varianza muestral debido a la reducción del tamaño de la muestra.

En décadas anteriores era habitual, a la hora de analizar datos, ignorar aquellos registros que poseían datos faltantes. Por un lado las estimaciones pueden estar sesgadas, ya que la eliminación de estos registros, supone que la no-respuesta se distribuye de forma aleatoria entre los distintos tipos de entrevistados. En el mejor de los casos, aquel en el que la no-respuesta se distribuye de forma aleatoria, estamos perdiendo una cantidad importante de información al eliminar los datos que estos individuos proporcionan a otras preguntas o proposiciones del cuestionario.

2.1 Métodos que emplean toda la Información disponible

Los métodos que emplean toda la información disponible consisten en considerar para los sucesivos análisis únicamente la información completa de las variables investigadas. Existen dos métodos que se comentan a continuación:

2.2.1 Eliminación por Filas

El método de eliminación por filas consiste en emplear solamente los registros que tengan respuesta en todas las variables de estudio, es decir solo para los entrevistados que contesten todas las preguntas o cuyos datos fueron íntegramente digitados. Las ventajas de este método son su simplicidad pero se desperdicia información que se conoce. [2]

Para ilustrar este método, se tiene una matriz de datos cuyas columnas son muestras tomadas de tres poblaciones todas ellas Poisson, independientes e idénticamente distribuidas con parámetro conocido $\lambda = 5$, $\mathbf{X} \in \mathrm{M}_{5x3}$, $i = 1,2,3,4,5$ y $j = 1,2,3$ y se supone que tiene el 13% de datos faltantes, es decir dos datos, los que recayeron en las variables X_2 y X_3 y son: el $X_{2,2}$=4 y $X_{4,3}$=7. Nótese que el 13% de datos faltantes en la matriz, constituye el 20% de datos faltantes en la columna que corresponde a X_2 y 20% de datos faltantes en la columna X_3. (Ver Tabla I)

Tabla I

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias independientes con distribución Poisson $\lambda = 5$

Tamaño de muestra n=5

X_1	X_2	X_3
8	4	6
4	**4**	5
3	5	6
1	7	**7**
6	5	2

El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \end{pmatrix} = \begin{pmatrix} 4.400 \\ 5.000 \\ 5.200 \end{pmatrix}$$

Como tenemos dos datos faltantes entonces se procede a prescindir de las dos filas que contienen los mismos y la matriz de datos ahora de datos resultante es (Ver Tabla II)

Tabla II

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias independientes con distribución Poisson $\lambda = 5$

Mètodo de Eliminaciòn por Filas

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

X_1	X_2	X_3
8	4	6
3	5	6
6	5	2

El vector de medias para las tres filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \end{pmatrix} = \begin{pmatrix} 5.667 \\ 4.667 \\ 4.667 \end{pmatrix}$$

Como era de esperarse el vector de medias de los datos originales y de los datos con filas eliminadas no coincide.

Ahora analicemos el efecto que causa en la matriz de varianzas y covarianzas, la eliminación de dos filas, con un tamaño de muestra *n*= 5 (Ver Cuadro I)

CUADRO I

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución

Poisson $\lambda = 5$

Método de eliminación por Filas

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3
X_1	7.300		
X_2	-2.500	1.500	
X_3	-2.350	0.750	3.700

Matriz de Varianzas y Covarianzas (Dos Filas Eliminadas)

	X_1	X_2	X_3
X_1	6.333		
X_2	-1.167	0.333	
X_3	-0.667	-0.667	5.333

Analizando el Cuadro I se puede apreciar que las covarianzas entre las variables disminuyeron, en la matriz con dos filas eliminadas, tal es el caso de la covarianza entre X_1 y X_3, la que disminuye de 0.750 a 0.667.

2.2.2 Eliminación por Pares

El método de eliminación por pares emplea todas las observaciones que tienen valores válidos para las variables de interés en cada momento, es decir usa todas las observaciones disponibles cuando calculamos $\bar{X}$ y todos los pares disponibles de valores en el cálculo de la matriz de correlación **R** y la matriz de covarianzas **S.** [2]

Para ilustrar consideraremos la siguiente matriz de datos:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & - & X_{23} \\ X_{31} & X_{32} & X_{33} \\ X_{41} & X_{42} & - \\ X_{51} & X_{52} & X_{53} \end{bmatrix} \quad X \in M_{5x3}$$

Para obtener $\overline{\mathrm{X}}_1$ se tienen cinco observaciones; para $\overline{\mathrm{X}}_2$ y $\overline{\mathrm{X}}_3$ se tienen cuatro observaciones disponibles. Para S_{12} y S_{13}, hay cuatro pares de

observaciones; para s_{23}, solo tres pares están disponibles.

A simple vista, esta forma de aproximarse al problema es atractiva porque usa toda la información disponible, pero el procedimiento generalmente no se recomienda ya que para el estudio de la correlación o covarianza entre las distintas variables el número de elementos variará según el número de registros que no tengan valores faltantes en dichas variables.

Se ilustra este método utilizando los mismos datos del ejemplo anterior, es decir, una matriz de datos cuyas columnas son muestras tomadas de tres poblaciones todas ellas Poisson, y se supone que tiene el 13% de datos faltantes, dos datos, los que recayeron en las variables X_2 y X_3 y son: el $X_{2,2}$=4 y $X_{4,3}$=7.

Tabla III

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de Eliminación por Pares

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

X_1	X_2	X_3
8	4	6
4	**4**	5
3	5	6
1	7	**7**
6	5	2

Entonces para obtener $\overline{X}_1$ se tienen cinco observaciones, en cambio para $\overline{X}_2$ y $\overline{X}_3$ se tienen solo cuatro observaciones. Para S_{12} y S_{13}, hay cuatro pares de observaciones; para S_{23}, solo tres pares están disponibles y estos son:

Para S_{12} los pares de observaciones disponibles son: (8,4),(3,5),(1,7) y (6,5), ya que aquí se elimina un par de observaciones. (Ver Cuadro II)

CUADRO II

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de eliminación por Filas

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Pares de observaciones disponibles para s_{12}

X_1	X_2
8	4
3	5
1	7
6	5

Matriz de Varianzas y Covarianzas para s_{12}

Variables	X_1	X_2
X_1	9.670	
X_2	-3.500	1.580

Para S_{13} los pares de observaciones disponibles son: (8,6),(4,5),(3,6) y (6,2)

CUADRO III

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de eliminación por Filas

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Pares de observaciones disponibles para s_{13}

X_1	X_2
8	6
4	5
3	6
6	2

Matriz de Varianzas y Covarianzas para s_{13}

Variables	X_1	X_2
X_1	4.920	
X_2	-0.580	3.580

Para S_{23} los pares de observaciones disponibles son: (4,6),(5,6) y (5,2)

CUADRO IV

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de eliminación por Filas

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Pares de observaciones disponibles para s_{23}

X_1	X_2
4	6
5	6
5	2

Matriz de Varianzas y Covarianzas para s_{23}

Variables	X_1	X_2
X_1	0.330	
X_2	-0.670	5.330

Donde la matriz de correlaciones queda de la siguiente forma:

CUADRO V

Efectos de la imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de eliminación por Pares

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas

Variables	X_1	X_2	X_3
X_1	1		
X_2	-3.500	1	
X_3	-0.580	-0.670	1

Este método tiene la desventaja de no poder asegurar que la matriz de correlaciones sea definida positiva, condición indispensable para invertir la matriz de correlaciones. Esta situación es debido a que se emplean distintas submuestras para el cálculo de las distintas correlaciones.

3. Modelos Estocásticos a utilizarse para Imputación de Datos

Los números "pseudo aleatorios" son la base en la construcción de los modelos de simulación donde hay presencia de variables estocásticas, ya que estos permiten el funcionamiento de las abstracciones con los que un fenómeno que no se puede construir físicamente, sea numéricamente construìdo o recreado. Existe un gran número de métodos que permiten la generación de números aleatorios entre 0 y 1, la importancia del método a utilizar radica en los números que genera, ya que estos números deben cumplir ciertas características para que sean válidos. Dichas características son:

- Ser uniformemente distribuidos.
- Ser estocàsticamente independientes lo cual significa que si X_1 y X_2 son dos variables aleatorias**,** X_1 y X_2, son independientes si y sólo si $f_{12}(X_1, X_2) = f_1(X_1) f_2(X_2)$; siendo f_{12} la distribución conjunta de X_1 y X_2 y además f_1 y f_2 las marginales de X_1 y X_2 respectivamente.
- Además es recomendable que los períodos del generador sean "largos" es decir sin repetición dentro de una longitud determinada de la sucesión de valores generados.

3.1. Generadores Congruenciales Lineales

La generación de números pseudos aleatorios se realiza a través de una "relación de recurrencia", es decir para una sucesión $X_0, X_1, ..., X_n$, es una expresión que define a cada término X_n, en función de uno o más de los términos que le preceden. Los valores de los términos necesarios para empezar a calcular se llaman condiciones iniciales. Se han propuesto varios esquemas como los *métodos congruenciales*: *congruencial mixto* y *congruencial multiplicativo.*

3.1.1 Método Congruencial Mixto

El *Método Congruencial Mixto* genera una sucesión de números pseudo aleatorios en la cual el sucesor X_{n+1} del número pseudo aleatorio X_n es determinado justo a partir de X_n. Particularmente para el caso del generador congruencial mixto la relación de recurrencia es la siguiente:

$$X_{n+1} = (aX_n + c) \bmod m$$

Donde:

$X_0 > 0$: representa la semilla y es un valor que elige el investigador;

$a > 0$: se denomina multiplicador;

$c > 0$: es una constante aditiva la que se denomina incremento; m es el "módulo", siendo;

$m > X_0$, $m > a$ y además $m > c$

Esta "relación de recurrencia" nos dice que X_{n+1} es el residuo de dividir $aX_n + c$ para el módulo. Es decir que los valores posibles de X_{n+1} son $0, 1, 2, 3, ..., m-1$, tal que, m representa el número posible de valores diferentes que pueden ser generados.

Para ilustrar la generación de números pseudoaleatorios por medio de este método, suponga que se tiene un generador en el cual los valores de sus parámetros son: $a = 5$, $c = 7$, $X_0 = 4$ y $m = 8$.

Como se puede apreciar en la Tabla IV el "período del generador" es ocho, esto es la sucesión se repite una vez que se obtuvo el octavo número generado.

TABLA IV

Efectos de la Imputación en el análisis de datos multivariados

Método Congruencial Mixto

Números pseudos aleatorios del generador

$X_{n+1} = (5X_n + 7) \bmod 8$

n	X_n	$(5X_n+7)/8$	X_{n+1}	Números Uniformes
0	4	3+3/8	3	0.375
1	3	2+6/8	6	0.750
2	6	4+5/8	5	0.625
3	5	4+0/8	0	0.000
4	0	0+7/8	7	0.875
5	7	5+2/8	2	0.250
6	2	2+1/8	1	0.125
7	**1**	**1+4/8**	**4**	**0.500**
8	4	3+3/8	3	0.375
9	3	2+6/8	6	0.750
10	6	4+5/8	5	0.625
11	5	4+0/8	0	0.000
12	0	0+7/8	7	0.875

Al analizar este ejemplo se podría pensar que el período de todo generador es siempre igual a *m*. Sin embargo, esto no es verdad ya que el período del generador depende de los valores asignados a los

parámetros a, c, X_0 y m, es decir, se requiere seleccionar valores adecuados para estos parámetros con el fin de que el generador tenga "período largo".

Con el fin de ilustrar el caso que se presenta cuando el *período del generador* es menor que m, suponga que se tiene un caso en el cual los valores de los parámetros son: $a = X_0 = c = 7$ y $m = 10$. Para estos valores, la sucesión de números pseudo aleatorios y uniformes son mostrados en la Tabla V.

Se puede apreciar que el período del generador es cuatro, lo cual deja claro que una selección inadecuada de los valores de los parámetros del generador, puede conducirnos a obtener períodos indeseables.

TABLA V

Efectos de la Imputación en el análisis de datos multivariados
Método Congruencial Mixto
Números pseudoaleatorios del generador
$X_{n+1} = (7X_n + 7) \bmod 10$

n	X_n	$(7X_n+7)/10$	X_{n+1}	Números Uniformes
0	7	5+6/10	6	0.600
1	6	4+9/10	9	0.900
2	9	7+0/10	0	0.000
3	**0**	**0+7/10**	**7**	**0.700**
4	7	5+6/10	6	0.600
5	6	4+9/10	9	0.900
6	9	7+0/10	0	0.000

El valor apropiado del módulo *m* debe ser el número entero más grande que la computadora acepte, el multiplicador a debe ser un entero impar no divisible para 3 ó 5, la constante aditiva c, puede ser cualquier constante y el valor de la semilla X_0, es irrelevante, para el generador congruencial mixto, es decir, el valor de este parámetro resulta tener poca o ninguna influencia sobre las propiedades estadísticas de las sucesiones.

3.1.2 Método Congruencial Multiplicativo

El *Método Congruencial Multiplicativo* al igual que el congruencial mixto genera una sucesión de números pseudos aleatorios en la cual el sucesor X_{n+1} del número pseudo aleatorio X_n es determinado justo a partir de X_n, de acuerdo a la siguiente relación de recurrencia:

$$X_{n+1} = aX_n \bmod m$$

Para ilustrar la obtención del período de un generador utilizando el *Método Congruencial Multiplicativo*, suponga que se tiene un generador con los siguientes parámetros: $a = 5$, $X_0 = 5$ y *m*=32. Estos valores se muestran en la Tabla VI.

TABLA VI

Efectos de la imputación en el análisis de datos multivariados
Método Congruencial Multiplicativo
Números pseudoaleatorios del generador
$X_{n+1} = 5X_n \bmod 32$

n	X_n	$5X_n / 32$	X_{n+1}	Números Uniformes
0	5	0+25/32	25	0.781
1	25	3+29/32	29	0.906
2	29	4+17/32	17	0.531
3	17	2+21/32	21	0.656
4	21	3+9/32	9	0.281
5	9	1+13/32	13	0.406
6	13	2+1/32	1	0.031
7	**1**	0+5/32	**5**	**0.156**
8	5	0+25/32	25	0.781
9	25	3+29/32	29	0.906
10	29	4+17/32	17	0.531
11	17	2+21/32	21	0.656

Se puede apreciar en Tabla VI que el período del generador es ocho, esto es la sucesión se repite una vez que se obtuvo el octavo número generado.

3.2 Métodos de Generación de Variables Aleatorias No Uniformes

3.2.1 Método de la Transformada Inversa

El "Método de la Transformada Inversa" utiliza la distribución acumulada *F(x)* de una variable aleatoria *X* que se va a simular. Puesto que *F(x)* está definida en el intervalo (0,1), y que además *F(x)=x* para $x \in (0,1)$ se puede generar un número aleatorio uniforme *y* y tratar de determinar el valor de la variable aleatoria para la cual su distribución acumulada es igual a *y*. Recordemos que *F* es una función sobreyectiva e inyectiva y por tanto un isomorfismo, además $\lim_{x \to -\infty} F(x) = 0$ y $\lim_{x \to \infty} F(x) = 1$.

Para convertir a un valor *x*, tomado de una distribución específica, a partir de un valor uniforme, se deberá encontrar *y* en términos de *x*, a partir de:

$$F(x) = y$$
$$F^{-1}(F(x)) = F^{-1}(y)$$
$$x = F^{-1}(y)$$

Este método tiene la dificultad principal de que en algunas ocasiones es difícil encontrar la transformada

inversa. Sin embargo, si esta función inversa ya ha sido establecida, generando números aleatorios uniformes se podrán obtener valores de la variable aleatoria que sigan la distribución de probabilidad deseada.

4. Técnicas de Imputación Aplicables

4.1 Imputación de Datos

Se entiende por "imputación de datos" a la acción de reemplazar, con algún criterio, los datos faltantes esto es, aquellos que por una u otra razón no se encuentren presentes en una matriz de datos; para de esta forma obtener un conjunto de "datos completos" con los que se pretende mantener, en lo posible, las características de la población objetivo investigada.

En las últimas décadas, se han desarrollado gran variedad de métodos de imputación para enfrentar el problema de datos faltantes y obtener una "matriz de datos completa".

4.1.1 Métodos de Imputación

Entre los métodos de imputación más difundidos y que son los que formarán parte de esta investigación están: asignar la *media aritmética* de los datos incompletos al o los valores faltantes y predecir el valor ausente mediante un *modelo de regresión*.

4.1.1.1 Imputación por la media muestral

El método de imputación por la media muestral, denominado también método de Wilks (1932), es muy sencillo de aplicar y útil para variables continuas aún cuando presentan inconvenientes estadísticos; consiste en la asignación en la matriz de datos del valor promedio de los datos existentes en la correspondiente columna, a todos los valores que "le faltan" a la matriz de datos $\mathbf{X} = (X_{ij})$, variable por variable. Supongamos que para una variable X_j tenemos registrados r de los n valores investigados y $(n-r)$ "datos faltantes", por lo que para los $(n-r)$ datos, los valores a ser imputados en la variable X_j se determinan así:

$$X_{(imp)j} = \frac{\sum_{i=1}^{r} X_{i(obs)j}}{r}$$

Siendo $X_{(imp)j}$ el valor que se coloca, "o imputa", en la variable con datos faltantes.
Sin embargo, este método tiene como desventaja que modifica la distribución de la variable, disminuyendo la variabilidad de los datos; de igual manera en el caso de realizar análisis multivariados se distorsiona la matriz de varianzas y covarianzas entre las variables observadas. Es decir, este método no conserva la relación entre las variables ni la distribución de frecuencias original. [2]

A continuación se ilustra este método:

Se tiene una matriz de datos cuyas columnas son muestras tomadas de cuatro poblaciones todas ellas Poisson, y que son estocàsticamente independientes entre sí, la primera variable tiene parámetro $\lambda = 2$, la segunda variable $\lambda = 4$, la tercera variable $\lambda = 5$ y la cuarta variable $\lambda = 7$, esto es:

$$f(X_1) = P(X_1 = x_1) = \frac{2^{x_1} e^{-2}}{x_1!}, \quad x_1 = 0,1,2,...$$

$$f(X_2) = P(X_2 = x_2) = \frac{4^{x_2} e^{-4}}{x_2!}, \quad x_2 = 0,1,2,..$$

$$f(X_3) = P(X_3 = x_3) = \frac{5^{x_3} e^{-5}}{x_3!}, \quad x_3 = 0,1,2,..$$

$$f(X_4) = P(X_4 = x_4) = \frac{7^{x_4} e^{-7}}{x_4!}, \quad x_4 = 0,1,2,..$$

Caso: Falta un dato en solo una variable

Se supone que la variable aleatoria X_4 que proviene de una distribución Poisson con $\lambda = 7$, tiene un valor faltante, el X_{74}, que realmente es igual a 14 (Ver Tabla VII). Nótese que, un dato faltante representa, en este caso, el 3% de datos faltantes en la matriz de datos.

TABLA VII

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias independientes con distribución Poisson

Tamaño de muestra n=10, 3% de datos faltantes en la matriz

X_1	X_2	X_3	X_4
5	4	3	6
1	7	1	6
2	6	8	10
2	5	3	2
4	6	4	9
3	5	6	12
2	3	4	**14**
0	3	5	9
3	3	2	6
2	4	11	7

El valor de la media aritmética de X_4, con el dato faltante es $\bar{X}_4 = \frac{6+6+10+2+9+12+9+6+7}{9} = 7.444$, entonces reemplazamos en $X_{74} = 7.444$, así calculamos nuevamente la media aritmética y la varianza con el dato imputado. El vector de medias de los datos originales es:

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \end{pmatrix} = \begin{pmatrix} 2.400 \\ 4.600 \\ 4.700 \\ 8.100 \end{pmatrix}$$

Mientras que el vector de medias con un dato completado es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \end{pmatrix} = \begin{pmatrix} 2.400 \\ 4.600 \\ 4.700 \\ 7.444 \end{pmatrix}$$

Analicemos ahora el efecto de esta imputación en la matriz de varianzas y covarianzas, comparando la matriz original con la matriz con 3% de datos completados mediante imputación por la media (Ver Cuadro VI)

Por medio del Cuadro VI podemos apreciar las varianzas y covarianzas entre las variables, utilizando la matriz de datos originales, las variables X_3 y X_4, muestran la mayor covarianza (3.033), seguida por la covarianza entre X_2 y X_4 (-0.844). También se aprecia un valor "grande" en la varianza de la variable X_4 (11.878), por ende valores de esta variable tienden a distribuirse lejos de la media, mientras que las variables X_1 y X_3 tienen la misma varianza (2.044). En la matriz de varianzas y covarianzas de los datos con imputaciòn por la media se nota una disminución en la varianza de la variable X_4, comparándola con la matriz de datos original; esto ocurre debido a que se inserta el valor de la media en los datos faltantes de esa variable y por ende los datos están menos dispersos.

CUADRO VI

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson

Método de Imputación por Media

Tamaño de muestra n=10 y 3% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

Variables	X_1	X_2	X_3	X_4
X_1	2.044			
X_2	0.067	2.044		
X_3	-0.533	-0.356	8.900	
X_4	-0.267	-0.844	3.033	**11.878**

Matriz de Varianzas y Covarianzas (Un Dato Completado con Imputación en X_4)

Variables	X_1	X_2	X_3	X_4
X_1	2.044			
X_2	0.067	2.044		
X_3	-0.533	-0.356	8.900	
X_4	-0.267	-0.844	3.543	**7.580**

4.1.1.2 Imputación por Regresión

El método de Imputación por Regresión se realiza particionando la matriz **X** en dos conjuntos, uno que contiene todas las filas con "valores faltantes" y otro las filas con "valores completos".

Supongamos que X_{ij} es el único valor faltante en la entrada de la i-ésima fila $\mathbf{X} \in M_{nx\,p}$, luego usamos los datos en la sub-matriz con las (p-1) filas completas, X_j se retrocede en las otras variables para obtener la ecuación de predicción

$$\hat{Y}_j = b_0 + b_1 X_1 + \ldots + b_{j-1} X_{j-1} + b_{j+1} X_{j+1} + \ldots + b_p X_p$$

El cálculo de los coeficientes de la regresión es de la forma:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Luego las entradas no faltantes en la i-ésima fila son como las variables de explicación en la ecuación de regresión para predecir el valor X_{ij}. El método de regresión utilizado para imputar datos fue propuesto primero por Buck (1960).

El método de regresión puede ser mejorado por iteración, es decir, primero se estiman todas las "entradas" faltantes en la matriz de datos usando regresión, después se llena los espacios en las "entradas" faltantes, luego se utiliza la matriz de datos así completada para obtener la nueva ecuación de predicción. [2]

Se utiliza los nuevos datos de la matriz para obtener la ecuación revisada de predicción y los nuevos valores $\hat{X}_{ij}$ y se continúa el proceso hasta que los valores de predicción se estabilicen.

Si las variables tienen demasiadas filas con datos faltantes, para utilizar el algoritmo de regresión en primera instancia, se puede usar el método de imputación por la media y luego usar regresión en las siguientes iteraciones.

A continuación se ilustra esta técnica:

Se tiene una matriz de datos cuyas columnas son muestras tomadas de tres poblaciones todas ellas Normal, y que son dependientes, donde cada columna tiene parámetros μ y σ^2 conocidos, $\mathbf{X} \in \mathrm{M}_{10x3}$, i= 1,2,....10 y j= 1,2,3.

Caso: Faltan dos datos, uno en X_2 y uno en X_3

TABLA VIII

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Tamaño de muestra n=10, 7% de datos faltantes en la matriz

X_1	X_2	X_3
35.011	**3.500**	2.801
35.002	4.901	**2.702**
40.021	30.000	4.382
10.101	2.802	3.211
6.003	2.701	2.732
20.000	2.821	2.810
35.000	4.640	2.881
35.100	10.921	2.902
35.100	8.010	3.283
30.002	1.611	3.201

1° Paso

Particionamos la matriz de datos en dos partes:

TABLA IX

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Tamaño de muestra n=10, 7% de datos faltantes en la matriz

Matriz particionada

X_1	X_2	X_3
35.011	**3.500**	2.801
35.002	4.901	**2.702**
40.021	30.000	4.382
10.101	2.802	3.211
6.003	2.701	2.732
20.000	2.821	2.810
35.000	4.640	2.881
35.100	10.921	2.902
35.100	8.010	3.283
30.002	1.611	3.201

Una parte de la matriz tiene filas con valores completos y la otra parte tiene filas con valores faltantes (Ver Tabla IX)

2° Paso

Utilizamos los datos de la sub-matriz con las filas completas para hacer la predicción. Las unidades con filas completas serán las variables independientes.

Primero X_1 y X_3 son las variables independientes que van a explicar a X_2; para las observaciones tercera a la décima, utilizando la ecuación de regresión $\hat{X}_2 = b_0 + b_1 X_1 + b_3 X_3$;

Los valores de los betas se los evalúa en las dos entradas de valores completos en la primera fila (X_1=35.011 , X_3=2.801)

$$\hat{X}_2 = b_0 + b_1(35.011) + b_3(2.801)$$

$$\hat{X}_2 = (-39.840) + (0.154)(35.011) + (13.801)(2.801)$$

$$\hat{X}_2 = 4.208$$

Similarmente hacemos la siguiente regresión pero ahora X_1 y X_2 son las variables independientes que van a explicar a X_3, donde $b_0 = 2.814, b_1 = -0.001, b_3 = 0.051$, los que se evalúan en las dos entradas de valores completos en la segunda fila (X_1=35.002, X_3=4.901)

$$\hat{X}_3 = b_0 + b_1(35.002) + b_2(4.901)$$

$$\hat{X}_3 = (2.814) - (0.001)(35.002) + (0.051)(4.901)$$

$$\hat{X}_3 = 3.029$$

3° Paso

Ahora insertamos estos estimadores, 4.208 en X_{12} y 3.099 en X_{23}, en los valores faltantes y calculamos la ecuación de regresión basada en las diez observaciones.

$$\hat{X}_2 = (-40.526) + (0.138)(35.011) + (14.063)(2.801)$$

$$\hat{X}_2 = 3.696$$

$$\hat{X}_3 = (2.828) - (0.003)(35.002) + (0.052)(4.901)$$

$$\hat{X}_3 = 2.978$$

4° Paso

Nuevamente insertamos los estimadores calculados, 3.696 en X_{12} y 2.978 en X_{23} y calculamos la nueva ecuación para obtener los valores de predicción.

$$\hat{X}_2 = (-40.680) + (0.139)(35.011) + (14.114)(2.801)$$

$$\hat{X}_2 = 3.720$$

$$\hat{X}_3 = (2.831) + (-0.003)(35.002) + (0.052)(4.901)$$

$$\hat{X}_3 = 2.981$$

Aquí hay un cambio en las siguientes iteraciones. Estos valores ($\hat{X}_2 = 3.720$, $\hat{X}_3 = 2.981$) tienden a los

verdaderos valores (X_2=3.500 y X_3=2.702) que inicialmente la regresión estimó así ($\hat{X}_2 = 4.208$ y $\hat{X}_3 = 3.099$). Además si se realizaba la imputación por la media los valores de X_2 y X_3 serían $\overline{X}_2 = 7.601$ y $\overline{X}_3 = 3.134$.

5. Simulación bajo distintas condiciones univariadas y multivariadas

Se presentan y analizan los resultados obtenidos al comparar los métodos de imputación utilizando diferentes tamaños de muestras, así como distintas distribuciones continuas y discretas tales como: normal, poisson y exponencial. El análisis se lo realiza para variables aleatorias conjuntas dependientes e independientes.

5.1 Matriz de datos con variables aleatorias independientes

Se tiene una matriz de datos cuyas columnas son muestras tomadas de cinco poblaciones todas ellas Normal, independientes e idénticamente distribuidas, con parámetros μ=5 y σ^2=1, $\mathbf{X} \in \mathrm{M}_{30x5}$, i= 1,2,....30 y j= 1,2,3,4,5 y se supone que tiene el 5% de datos faltantes, es decir tres datos, los que recayeron en la variable X_1 y son: el $X_{10,1}$=4.168, $X_{14,1}$=6.624 y el $X_{25,1}$=6.290. Nótese que el 5% de datos faltantes en la matriz, constituye 10% de datos faltantes en la columna que corresponde a X_1
El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.205 \\ 4.970 \\ 4.984 \\ 4.608 \\ 4.955 \end{pmatrix}$$

Método de Imputación por la Media y Regresión

Por medio del Método de Imputación por la Media, se procede a calcular la media aritmética de la variable X_1 con los tres datos faltantes, cuyo valor es 5.151, entonces reemplazamos en $X_{10,1}$, $X_{14,1}$ y en $X_{25,1}$ y por medio del Método de Imputación por Regresión, el cálculo de los valores faltantes se realiza por medio de la ecuación de predicción y el cálculo de los coeficientes de la misma es de la forma $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.
En la Tabla X se realiza una comparación entre el valor real y el valor con imputación por la media y regresión.

Tabla X
Efectos de la Imputación en el análisis de datos multivariados
Variables aleatorias independientes con distribución Normal (5,1)
Comparación de los Métodos de Imputación
Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

10% de datos completados en X_1 por la Media

Dato Observado	Imputación por la Media	*Error* | *Dato Observado* – Dato con Imputación |
4.168	5.151	0.983
6.624	5.151	1.473
6.290	5.151	1.139

10% de datos completados en X_1 por Regresión

Dato Observado	Imputación por Regresión	*Error* | *Dato Observado* – Dato con Imputación |
4.168	5.245	1.077
6.624	5.871	0.753
6.290	5.726	0.564

La diferencia en valor absoluto entre el dato observado de cada variable es menor en el "Método de Imputación por Regresión", con excepción del primer valor donde error por medio del Método de Imputación por Media es menor (0.983), pero en general los valores no tienden al dato observado.

5.1 Matriz de datos con variables aleatorias dependientes

Se tiene una matriz de datos cuyas columnas son muestras tomadas de cinco poblaciones todas ellas Normal, dependientes e idénticamente distribuidas, con parámetros μ=10 y σ^2=1, $\mathbf{X} \in \mathrm{M}_{50x5}$, *i= 1,2,....50* y *j= 1,2,3,4,5* y se supone que tiene el 5% de datos faltantes, es decir trece datos, los que recayeron en la variable X_3 y son: el $X_{2,3}$=9.010, $X_{5,3}$=11.221, $X_{6,3}$=10.102, $X_{9,3}$=9.927, $X_{11,3}$=10.718, $X_{17,3}$=11.504, $X_{21,3}$=12.263, $X_{23,3}$=10.329, $X_{29,3}$=10.655, $X_{32,3}$=9.547, $X_{37,3}$=9.509, $X_{41,3}$=9.189 y el $X_{46,3}$=9.549 . Nótese que el 5% de datos faltantes en la matriz, constituye 26% de datos faltantes en la columna que corresponde a X_3.

El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 9.997 \\ 10.012 \\ 10.214 \\ 10.137 \\ 10.188 \end{pmatrix}$$

Método de Imputación por la Media y Regresión

Por medio del *Método de Imputación por Media*, se procede a calcular la media aritmética de la variable X_3 con los trece datos faltantes, cuyo valor es 10.194, entonces reemplazamos en $X_{2,3}$, $X_{5,3}$, $X_{6,3}$, $X_{9,3}$, $X_{11,3}$, $X_{17,3}$, $X_{21,3}$, $X_{23,3}$, $X_{29,3}$, $X_{32,3}$, $X_{37,3}$, $X_{41,3}$ y en $X_{46,3}$, así como también se calculan las ecuaciones de predicción para el Método de Imputación por Regresión.

Se puede notar, por medio de la Tabla XI que la diferencia en valor absoluto entre el dato observado de cada variable y el resultado de predicción, es menor en el Método de Imputación por Regresión, así como también que los valores tienden al dato observado.

Tabla XI

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

Comparación de los Métodos de Imputación

Tamaño de muestra n=50 y 5% de datos faltantes en la matriz

26% de datos completados en X_3 por la Media

Dato Observado	Resultado de Imputación por Media	*Error* |*Dato Observado* – Resultado de Imputación por Media|
9.010	10.194	1,184
11.221	10.194	1,027
10.102	10.194	0,092
9.927	10.194	0,267
10.718	10.194	0,524
11.504	10.194	1,310
12.263	10.194	2,069
10.329	10.194	0,135
10.655	10.194	0,461
9.547	10.194	0,647
9.509	10.194	0,685
9.189	10.194	1,005
9.549	10.194	0,645

26% de datos completados en X_3 por Regresión

Dato Observado	Resultado de Predicción	*Error* | *Dato Observado* – Resultado de Predicción |
9.010	9.110	0,100
11.221	11.215	0,006
10.102	10.112	0,010
9.927	9.931	0,004
10.718	10.709	0,009
11.504	11.510	0,006
12.263	12.253	0,010
10.329	10.333	0,004
10.655	10.652	0,003
9.547	9.545	0,002
9.509	9.507	0,002
9.189	9.181	0,008
9.549	9.539	0,010

6. Conclusiones

a) Si se trabaja con una matriz de datos cuyas columnas son muestra tomadas de poblaciones normales, independientes e idénticamente distribuidas, con un tamaño de muestra n=30 y 2% de datos faltantes, el *Método de Eliminación por Filas*, distorsiona el vector de medias de la matriz de datos originales, puesto que se eliminan filas para calcularlo, pero esta distorsión no afecta mayormente a la *matriz de varianzas y covarianzas* y *de correlaciones,* lo mismo sucede con la distribución poisson y exponencial.

b) Si se tiene una matriz de datos cuyas columnas son muestras tomadas de poblaciones Poisson, independientes e idénticamente distribuidas, con un tamaño de muestra mayor o igual a 30 y la cantidad de filas eliminadas es mayor o igual al 5%, la *matriz de varianzas y covarianzas* y *de correlaciones*, se ve afectada puesto que las covarianzas y correlaciones entre las variables varían considerablemente; lo mismo sucede con distribuciones normales y exponenciales.

c) Cuando se trabaja con matrices de datos cuyas columnas son muestras tomadas de poblaciones normales, dependientes e idénticamente distribuìdas con un tamaño de muestra mayor o igual a 50 y la cantidad de datos faltantes es del 5%, el método de eliminación por filas no afecta mayormente a la *matriz de varianzas y covarianzas* y *de correlaciones* ya que las variables están correlacionadas.

d) El *Método de Imputación por Regresión* es preferible al *Método de Imputación por Media*, cuando se trabaja con matrices de datos con variables aleatorias dependientes.

7. Referencias Bibliográficas

[1] MENDENHALL, W., WACKERLY, D. Y SCHEAFFER, R. (1994), "Estadística Matemática con Aplicaciones", Iberoamérica, México.

[2] RENCHER, A. (1998), "Multivariate Statistical Analysis and Applications" New York: Wiley series in Probability and Statistics.

[3] MARTÍNEZ, W.; MARTÍNEZ, A. (2002) "Computational Statistics Handbook with Matlab", Chapman & Hall/CRC, Boca Raton, United Sates of America.

[4] RIAL, A., VARELA, J., & ROJAS, A. (2001) "Depuración y Análisis Preliminares de Datos en SPSS", Sistemas Informatizados para la investigación del comportamiento, Edición RA-MA, Madrid-España.

_____
M. Sc. Gaudencio Zurita
Director de Tesis de Grado