ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

Ingeniería en Estadística Informática

"Efectos de la Imputación en el Análisis de Datos Multivariados"

TESIS DE GRADO

Previa la obtención del título de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

Presentada por:

Marcia Gabriela Cuenca Calle

GUAYAQUIL – ECUADOR

AÑO

2006

AGRADECIMIENTO

A Dios y la Virgen Santísima, por haberme permitido llegar hasta aquí. A mis padres y hermana, por la ayuda incondicional. A todos mis amigos: Eduardo, Emma, Patricio, Freddy, Juan, David, Mónica, Jorge, Fico, Fátima, Evelyn, etc. que con su ayuda y apoyo constante, han estado siempre presente.

A todos ellos,

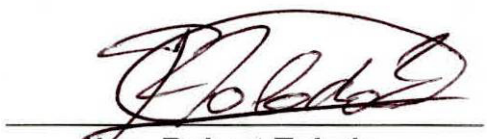
Muchas Gracias

DEDICATORIA

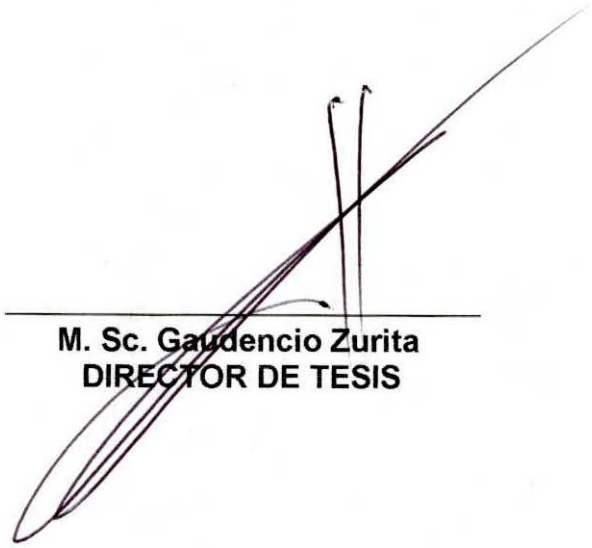
A los seres que me enseñaron que la responsabilidad, el esfuerzo y la perseverancia son los únicos medios para alcanzar el éxito.

A mis padres y hermana, *ERNESTINA, MANRIQUE y PAOLA.*

TRIBUNAL DE GRADUACIÓN



Ing. Robert Toledo
SUB-DIRECTOR DEL ICM



M. Sc. Gaudencio Zurita
DIRECTOR DE TESIS

Ing. Erwin Delgado
VOCAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, me corresponde exclusivamente, y el patrimonio intelectual de la misma a la Escuela Superior Politécnica del Litoral"

(*Reglamento de Graduación de la ESPOL*)

Marcia Gabriela Cuenca Calle

RESUMEN

El presente trabajo consiste en un estudio estadístico acerca de los Efectos de la Imputación en el Análisis de Datos Multivariados, basados en muestras con variables aleatorias dependientes e independientes de diferentes tamaños y distribuciones, así como también el análisis de un caso real.

La tesis está conformada por cuatro capítulos más las conclusiones y recomendaciones. El primer capítulo describe los principios estadísticos relacionados con los Métodos de Imputación que son parte de esta investigación.

En el capítulo dos aborda las técnicas y principios científicos que permiten la generación de números aleatorios. El tercer capítulo ilustra las técnicas de imputación para el manejo de datos incompletos en una matriz de datos. En el siguiente capítulo se comparan los métodos de imputación por medio de simulaciones. Finalmente se muestran las conclusiones y recomendaciones basadas en los resultados obtenidos en este trabajo.

INDICE GENERAL

SIMBOLOGÍA

$X \in M_{nxp}$	Matriz de datos multivariada.
P	Población
Ω	Conjunto de todos lo resultados posibles del experimento
δ	Es el δ - álgebra de subconjuntos de Ω
$\Re$	Conjunto de los Números reales.
X	Variable Aleatoria
μ	Media Poblacional
σ^2	Varianza Poblacional
$M_X(t)$	Función Generadora de Momentos
$\binom{N}{n}$	Número de subconjuntos, de tamaño n, entre N objetos disponibles.
$\overline{X}$	Media Muestral.
$\mathrm{E}(\overline{X}) = \mu$	Estimador insesgado de la media poblacional μ
s^2	Varianza Muestral.
$(1-\alpha)$	Nivel de confianza al 100 %.
n	Tamaño de muestra.
N	Tamaño de la población.
ρ_{ik}	Coeficiente de correlación lineal entre las variables *i* y *k*.
σ	Desviación Estándar de la población.

$X \in \Re^p$	Vector Aleatorio *p*-variado.
$\sum$	Matriz de varianzas y covarianzas.
s_{ij}	Matriz muestral de varianzas y covarianzas.
σ_{ij}	Covarianza entre las variables *i* y *j*.
$\boldsymbol{D}_s$	Matriz Diagonal
f	Función de densidad.
$U(\alpha, \beta)$	Distribución Uniforme con parámetros α y β.
H_0	Hipótesis Nula
H_1	Hipótesis Alternativa.
χ^2	Estadístico Ji Cuadrado.
$D_{n,a}$	Estadístico K-S tabulado.
X_n	Número Pseudos aleatorio.
X_{n+1}	Sucesor de un número aleatorio.
F	Función Acumulada.
$X_{(imp)j}$	Valor que se coloca, "o imputa", en la variable con datos faltantes.
$\overline{X}_{n-1}$	Media para datos incompletos.

ÍNDICE DE TABLAS

Capítulo I

Capítulo II

Capítulo III

Capítulo IV

ÍNDICE DE GRÁFICOS

Capítulo II

ÍNDICE DE CUADROS

Capítulo I

INTRODUCCIÓN

La presente tesis tiene como objetivo efectuar en un estudio estadístico acerca de los Efectos de la Imputación en el Análisis de Datos Multivariados, el mismo que se basa en la generación de muestras con variables aleatorias dependientes e independientes de diferentes tamaños y distribuciones, así como también el análisis de un caso real.

El primer capítulo describe los principios estadísticos relacionados con los Métodos de Imputación que son parte de esta investigación, para esto presenta los conceptos relacionados con matrices de datos multivariados, y la "Pérdida de Datos" en una Investigación.

El capítulo dos aborda el tema de las técnicas y principios científicos que permiten la generación de números aleatorios, los mismos que son necesarios para la simulación de sistemas que se explican estocásticamente.

En el capítulo tres se ilustran las técnicas de imputación para el manejo de datos incompletos en una matriz de datos, para lo cual se define "Imputación de Datos" y los "Métodos de Imputación".

Por otro lado el capítulo cuatro, presenta y analiza los resultados obtenidos al comparar los métodos de imputación utilizando diferentes tamaños de muestras: 30, 50 y 100 así como distintas distribuciones continuas y discretas tales como: normal, poisson y exponencial.

En el último capítulo se muestran las conclusiones y las recomendaciones obtenidas del análisis de los resultados en este estudio.

CAPÍTULO I

1. LA PÉRDIDA DE DATOS EN UNA INVESTIGACIÓN

1.1 Introducción

El presente capítulo incluye los principios estadísticos relacionados con los Métodos de Imputación que serán parte de esta investigación. Para esto, se presenta, en la sección 1.2 los conceptos relacionados con matrices de datos multivariados, en la siguiente sección se muestra un resumen acerca de la "Pérdida de Datos" en una Investigación y por último se presentan los métodos que emplean toda la información disponible.

1.2 Matriz de Datos Multivariados

Una matriz es un arreglo rectangular de números reales, de n filas y p columnas que contiene información de una muestra aleatoria tomada de una población donde, por ejemplo, a n individuos se le realizan p preguntas. En el Cuadro 1.1, $\boldsymbol{X}$ es la matriz de datos y X_{ij} es el valor de la j-ésima variable investigada al i-ésimo individuo, es decir se miden p características a n individuos.

CUADRO 1.1

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos Multivariados

$$\mathbf{X}=\begin{bmatrix} X_{11} & X_{12} & X_{1p} \\ X_{21} & X_{22} & X_{2p} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ X_{n1} & X_{m2} & X_{np} \end{bmatrix} \quad ; \quad \mathbf{X} \in M_{nxp}$$

1.3 Variables aleatorias Univariadas y Bivariadas

1.3.1 Variables aleatorias univariadas

Sea (Ω, S) un espacio muestral, donde Ω es el conjunto de todos los resultados posibles del experimento y S es el conjunto potencia de Ω, X es una función de valor real definida sobre los elementos de (Ω, S), es decir que: $X:\Omega \rightarrow \Re$, entonces X es una *variable*

aleatoria siendo $\Re$ el conjunto de los Números Reales. Las variables aleatorias pueden ser continuas o discretas.

Variable Aleatoria Discreta

Una Variable Aleatoria Discreta X es, una variable aleatoria para la cual el número de valores $X(w), w \in \Omega$, que puede tomar, es finito o infinito numerable.

Variable Aleatoria Continua

Una Variable Aleatoria Continua X es, una variable aleatoria que toma valores $X(w), w \in \Omega$, en una escala continua, para dos variables cualesquiera siempre se puede encontrar un valor intermedio.

Población Objetivo

Se denomina Población Objetivo al conjunto de todos los elementos acerca de cuyas características deseamos hacer alguna investigación de tipo estadístico.

Población Investigada

La Población Investigada es el conjunto de entes pertenecientes a la Población Objetivo, disponibles al momento de efectuar la investigación, debido a que no siempre se puede acceder a todas

las unidades de investigación que conforman la población objetivo, ya sea por negativas a colaborar, ausencias o cualquier otro tipo de inaccesibilidad. Si todos los entes motivos de la investigación están disponibles, entonces la Población Objetivo es igual a la Población Investigada.

Valores Esperados y Varianza de una Variable Aleatoria

El valor esperado de una función g , dada en términos de X está denotada como $E[g(X)]$ y definida de la siguiente forma:

$$E[g(X)] = \int_{-\infty}^{\infty} g(X)\, f(x)dx \tag{1.1}$$

Si X es continuo y es tal que su función de densidad $f(x)$ es conocida, la media μ de la población o valor esperado de X es definida como:

$$E(X) = \mu = \int_{-\infty}^{\infty} X\, f(x)dx \tag{1.2}$$

Es simple demostrar que:

$$\text{a) } E(aX)=aE(X) \tag{1.3}$$

$$\text{b) } E[g(X)+h(X)]=E[g(X)]+E[h(X)] \tag{1.4}$$

La varianza poblacional $Var(X)$ es definida como:

(1.5)

$$Var(X) = \sigma^2 = E(X-\mu)^2$$

y la función generadora de momentos se define como

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tX} f(x)dx .$$

Utilizando (1.3) y (1.4), la varianza poblacional puede ser expresada como:

$$\sigma^2 = E(X^2) - \mu^2 \qquad (1.6)$$

La raíz cuadrada de la varianza poblacional es llamada como desviación estándar de la población.

Aparte de $\left.\frac{\delta M}{\delta t}\right|_{t=0} = E(X)$ y en general la $\left.\frac{\delta^r M}{\delta t^r}\right|_{t=0} = E(X^r)$

Si cada valor de X es multiplicado por una constante a, la varianza de la población de X se multiplica por a^2, es decir:

$$Var(aX)=a^2\sigma^2 \qquad (1.7)$$

Muestra

Una muestra $X_1, X_2, ..., X_n$, tomada de una población X, que es discreta, es aleatoria si y solo si, es escogida de tal forma que cada subconjunto de tamaño n en la población, tiene igual probabilidad

de constituir la muestra. La probabilidad de escoger una muestra de tamaño n de una población de tamaño N es $\frac{1}{\binom{N}{n}}$.

Una muestra $X_1, X_2, ..., X_n$, tomada de una población X, que es contìnua, es aleatoria, si y solo si $X_1, X_2, ..., X_n$ son variables aleatorias independientes e idénticamente distribuìdas.

La media aritmética $\overline{X}$ de una muestra aleatoria de tamaño n, $X_1, X_2, ..., X_n$ es definida por:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.8}$$

Si $X_1, X_2, ..., X_n$ es una muestra aleatoria de una población que tiene media μ y varianza σ^2, entonces la media de la muestra $\overline{X}$ es un estimador insesgado de la media poblacional μ, esto es:

$$E(\overline{X}) = \mu . \tag{1.9}$$

La media muestral tiene una propiedad similar a la que definimos en (1.3). Si el $Z_i = aX_i$ para $i = 1,2,3,...,n$, entonces $\overline{Z} = a\overline{X}$; veamos:

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n} aX_i = \frac{1}{n}a\sum_{i=1}^{n} X_i$$

$$\overline{Z} = a\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = a\overline{X}$$

$$\overline{Z} = a\overline{X} \qquad (1.10)$$

Para una muestra de n observaciones, la varianza muestral se define como:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \qquad (1.11)$$

La que también es igual a:

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1} \qquad (1.12)$$

Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una población con media μ y varianza σ^2, entonces la varianza muestral s^2 es un estimador insesgado de la varianza poblacional σ^2; esto es:

$$E(s^2) = \sigma^2 \qquad (1.13)$$

La cual se demuestra de la siguiente forma:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

$$
\begin{aligned}
E(s^2) &= E\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\bar{X}^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}E(X_i^2)-nE(\bar{X}^2)\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2+\mu^2)\right)-n\left(\frac{\sigma^2}{n}+\mu^2\right) \\
&= \frac{1}{n-1}\left(n\sigma^2+n\mu^2-\sigma^2-n\mu^2\right) \\
&= \frac{1}{n-1}\left(n\sigma^2-\sigma^2\right)=\frac{n-1}{n-1}\sigma^2=\sigma^2
\end{aligned}
$$

Similarmente, si definimos $Z_i=aX_i$, $i=1,2,\ldots,n$, entonces la varianza muestral de Z es dada por $s_Z^2=a^2s^2$, la cual demostraremos a continuación:

$$
\begin{aligned}
s_Z^2 &= \frac{\sum_{i=1}^{n}(Z_i-\bar{Z})^2}{n-1}=\frac{\sum_{i=1}^{n}(aX_i-a\bar{X})^2}{n-1} \\
&= \frac{\sum_{i=1}^{n}[a(X_i-\bar{X})]^2}{n-1}=\frac{a^2\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1} \\
&= a^2s^2
\end{aligned}
\qquad (1.14)
$$

1.3.2 Variables Aleatorias Bivariadas

Un vector aleatorio bivariado $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ surge cuando dos características X_1 y X_2 son medidas de manera simultánea en cada ente que se investiga.

La covarianza poblacional es definida como:

$$\operatorname{cov}(X_i, X_j) = \sigma_{ij} = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] \tag{1.15}$$

donde μ_i y μ_j son las medias de X_i y X_j respectivamente. Se puede demostrar que:

$$\sigma_{ij} = \mathrm{E}(X_i, X_j) - \mu_i \mu_j \tag{1.16}$$

Para una muestra (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) la covarianza muestral se define como:

$$s_{XY} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1} = \hat{\sigma}_{XY} \tag{1.17}$$

La que es equivalente a:

$$s_{XY} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{n-1} \tag{1.18}$$

La covarianza muestral s_{XY} es un estimador insesgado para la covarianza poblacional σ_{XY} es decir:

$$E(s_{XY}) = \sigma_{XY} \tag{1.19}$$

Puesto que la covarianza depende de la escala de la medida de X y Y, es difícil para comparar covarianzas entre diversos pares de variables. Por ejemplo, si cambiamos una medida de pulgadas a centímetros, la covarianza cambiará. Para encontrar una medida de la relación lineal que sea invariante a los cambios de escala, podemos estandardizar la covarianza dividiéndola para las desviaciones estándar de las dos variables. Esta covarianza estandardizada se llama usualmente coeficiente de correlación. La correlación poblacional de dos variables aleatorias X y Y es:

$$\rho_{XY} = corr(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E(X-\mu_X)^2}\sqrt{E(Y-\mu_Y)^2}} \tag{1.20}$$

Y la correlación muestral se da por:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{1.21}$$

El coeficiente de correlación poblacional y muestral es un valor entre -1 y 1.

1.3.3 Vectores Media y Matriz de Covarianza para Vectores Aleatorios

Supongamos que se tiene una muestra aleatoria multivariada de n vectores observados $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}$, tomada de una población

p-variada $\mathbf{X}$. Dos vectores $\mathbf{X_1}$ y $\mathbf{X_2}$ son independientes, si cada variable X_{1j} en $\mathbf{X_1}$ es independiente de cada variable X_{2j} en $\mathbf{X_2}$.

Ya que $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}$ constituye una muestra aleatoria, entonces sus n vectores son independientes.

Los n vectores observados son transpuestos y listados como filas en la matriz de datos $\mathbf{X} \in \Re^p$:

$$\mathrm{X} = \begin{pmatrix} \mathrm{X}_1^{\mathrm{T}} \\ \mathrm{X}_2^{\mathrm{T}} \\ . \\ . \\ . \\ \mathrm{X}_i^{\mathrm{T}} \\ . \\ . \\ . \\ \mathrm{X}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} \ldots X_{1j} \ldots X_{1p} \\ X_{21} & X_{22} \ldots X_{2j} \ldots X_{2p} \\ . & . \quad . \quad . \\ . & . \quad . \quad . \\ . & . \quad . \quad . \\ X_{i1} & X_{i2} \ldots X_{ij} \ldots X_{ip} \\ . & . \quad . \quad . \\ . & . \quad . \quad . \\ . & . \quad . \quad . \\ X_{n1} & X_{n2} \ldots X_{nj} \ldots X_{np} \end{pmatrix} \tag{1.22}$$

En la matriz $\mathbf{X}$, el primer subíndice representa unidades de investigación o individuos, y el segundo subíndice corresponde a las variables o características, donde en general $n > p$.

Si deseamos discutir ambas columnas y filas de $\mathbf{X}$, las columnas son denotadas de la siguiente manera:

$$\mathbf{X} = (\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \ldots, \mathbf{X}_{(p)}) \tag{1.23}$$

Así, por ejemplo $\mathbf{X}_2$ es el vector p-dimensional de las variables medidas en la segunda unidad investigada, mientras $\mathbf{X}_{(2)}$ es el *n*-vector de observaciones en la segunda variable.

El vector muestral es definido como:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ \cdot \\ \cdot \\ \cdot \\ \overline{\mathbf{X}}_p \end{pmatrix} \tag{1.24}$$

Así el promedio de los n vectores produce el promedio de cada variable.

Podemos calcular $\overline{\mathbf{X}}$ directamente de $\mathbf{X}$:

$$\overline{\mathbf{X}} = \frac{1}{n}\mathbf{X}'\mathbf{j} \text{ donde } \mathbf{j} \text{ es un vector } n\text{x}1 \text{ de unos } \quad \mathbf{j} = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \tag{1.25}$$

La media poblacional o valor esperado del vector aleatorio $\mathbf{X}$ es definido como el vector de valores esperados de p variables,

$$E(\mathbf{X}) = E\begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_p \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ . \\ . \\ . \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}, \tag{1.26}$$

donde $E(X_j) = \mu_j$. Ya que $E(\bar{X}_j) = \mu_j$, entonces:

$$\mathrm{E}(\bar{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_p \end{pmatrix} = \boldsymbol{\mu} \tag{1.27}$$

lo cual significa que $\bar{\mathbf{X}}$ es un estimador insesgado de $\boldsymbol{\mu}$.

La Matriz Muestral de Varianzas y Covarianzas es simétrica:

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdot & \cdot & \cdot & s_{1p} \\ s_{21} & s_{22} & \cdot & \cdot & \cdot & s_{2p} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ s_{p1} & s_{p2} & \cdot & \cdot & \cdot & s_{pp} \end{pmatrix}, \; s_{ij} = s_{ji} \tag{1.28}$$

Y por tanto diagonalizable ortogonalmente

La matriz de varianzas y covarianzas de la población es definida como:

$$\Sigma = \mathrm{E}[\ (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}\ ] \tag{1.29}$$

Donde resulta que Σ es una matriz cuadrada simétrica por lo tanto, diagonalizable ortogonalmente,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

El valor σ_{ij} es la covarianza entre X_i y X_j. Para el caso en que i sea igual a j, σ_{ij} es la varianza de la i-ésima variable X_i, σ_i^2, esto es $\sigma_{ii} = \sigma_i^2$.

1.3.4 Matriz de Correlación

La matriz de correlación poblacional está definida como:

$$\mathbf{P}_\rho = (\rho_{jk}) = \begin{pmatrix} 1 & \rho_{12}.....\rho_{1p} \\ \rho_{21} & 1.....\rho_{2p} \\ & \\ & \\ \rho_{p1} & \rho_{p2}.....1 \end{pmatrix} \tag{1.30}$$

donde $\rho_{jk} = \dfrac{\sigma_{jk}}{\sigma_j \sigma_k}$. El subíndice ρ en $\mathbf{P}_\rho$ es usado como recordatorio de que $\mathbf{P}$ es la versión mayúscula de ρ.

Si definimos $\mathbf{D}_\sigma = diag(\sigma_1, \sigma_2, ..., \sigma_p)$ será una matriz diagonal de la desviación de la población estándar análoga para D_S, luego:

$$\mathbf{P}_\rho = \mathbf{D}_\sigma^{-1} \sum \mathbf{D}_\sigma^{-1} \tag{1.31}$$

$$\sum = \mathbf{D}_\sigma \mathbf{P}_\rho \mathbf{D}_\sigma \tag{1.32}$$

Mientras $\overline{\mathbf{X}}$ y $\mathbf{S}$ son estimadores insesgados de μ y $\sum$, este no es el caso con $\mathbf{R}$.

Por (1.25) la correlación muestral entre las *j*-ésimas y *k*-ésimas variables está dada por:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}} = \frac{s_{jk}}{s_j s_k} \tag{1.33}$$

La matriz de correlación muestral es también una matriz de covarianzas definida como:

$$\mathbf{R} = (r_{jk}) = \begin{pmatrix} 1 & r_{12} \ldots\ldots r_{1p} \\ r_{21} & 1 \ldots\ldots r_{2p} \\ . & . \quad . \\ . & . \quad . \\ . & . \quad . \\ r_{p1} & r_{p2} \ldots\ldots 1 \end{pmatrix} \tag{1.34}$$

La cual es simétrica ya que $r_{jk} = r_{kj}$

$\mathbf{R}$ es una matriz de varianzas y covarianzas para datos estandarizados.

Para relacionar $\mathbf{R}$ (matriz de correlación muestral) y $\mathbf{S}$ (matriz de varianzas y covarianzas muestrales), se define la matriz diagonal:

$$\mathbf{D}_S = [diag(\mathbf{S})]^{1/2} = diag(s_1, s_2, \ldots, s_p) \tag{1.35}$$

Es posible probar que:

$$\mathbf{R} = \mathbf{D}_S^{-1}\, \mathbf{S}\, \mathbf{D}_S^{-1} \tag{1.36}$$

$$\mathbf{S} = \mathbf{D}_S\, \mathbf{R}\, \mathbf{D}_S \tag{1.37}$$

Si la matriz $\mathbf{X} = X_{ij}$ es estandarizada para $\mathbf{Z} = Z_{ij}$ donde $Z_{ij} = (X_{ij} - \bar{X})/s_j$ luego la matriz de covarianza de las zetas es igual a la matriz de correlación de las equis:

$$\mathbf{S}_Z = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} = \mathbf{R} \tag{1.38}$$

1.4 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 digitación, 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, y afectar a los valores de la matriz de varianzas y covarianzas y correlaciones.

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.

1.5 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:

1.5.1 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. [6]

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 $\lambda = 5$, $\mathbf{X} \in \mathrm{M}_{5\mathrm{X}3}$, *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 1.1)

Tabla 1.1

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

Elaborado por: G. Cuenca

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 1.2)

Tabla 1.2

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 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

Elaborado por: G. Cuenca

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.

CUADRO 1.2

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

Elaborado por: G. Cuenca

Analizando el Cuadro 1.2 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.

1.5.2 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 $\overline{X}$ y todos los pares disponibles de valores en el cálculo de la matriz de correlación $\mathbf{R}$ y la matriz de covarianzas $\mathbf{S}$. [6]

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 $\overline{X}_1$ se tienen cinco observaciones; para $\overline{X}_2$ y $\overline{X}_3$ 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, 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, 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 1.3
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

Elaborado por: G. Cuenca

Entonces para obtener X_1 se tienen cinco observaciones, en cambio para X_2 y 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 1.3)

CUADRO 1.3

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

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

Elaborado por: G. Cuenca

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

CUADRO 1.4

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

Pares de observaciones disponibles para s_{13}

X_1	X_3
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

Elaborado por: G. Cuenca

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

CUADRO 1.5

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

Pares de observaciones disponibles para s_{23}

X_1	X_3
4	6
5	6
5	2

Matriz de Varianzas y Covarianzas para s_{23}

Variables	X_1	X_3
X_1	0.330	
X_3	-0.670	5.330

Elaborado por: G. Cuenca

Donde la matriz de correlaciones es de la forma:

Tabla 1.4

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

Elaborado por: G. Cuenca

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.

CAPÍTULO II

2. MODELOS ESTOCÁSTICOS A UTILIZARSE PARA IMPUTACIÓN DE DATOS

2.1 Introducción

Este segundo capítulo aborda el tema de las técnicas y principios científicos que permiten la generación de números aleatorios, los mismos que son necesarios para la simulación de sistemas que se explican estocásticamente. Para esto, la sección 2.2 trata acerca de la Distribución Uniforme; la siguiente sección detalla los Métodos para la Generación de números aleatorios; después se describe un Método de Generación de Variables Aleatorias no Uniformes y por último se muestra las pruebas de hipótesis sobre los parámetros o la distribución de una población.

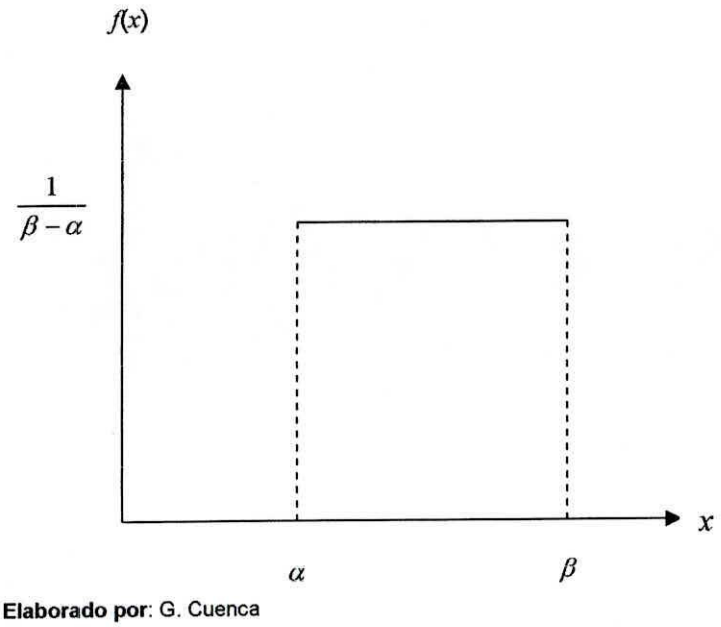
2.2 Distribución Uniforme

Una variable aleatoria X tiene una *distribución uniforme continua* con parámetros α y β si y solo sí su función de densidad f, está dada por

$$f(x)=\begin{cases}\dfrac{1}{\beta-\alpha}, & X\in(\alpha,\beta)\\ 0\ , & \textit{para el resto de } X\end{cases} \tag{2.1}$$

Los parámetros α y β de esta variable son constantes reales con $\alpha<\beta$, véase Gráfico 2.1.

GRÀFICO 2.1
Efectos de la Imputación en el análisis de datos multivariados
Densidad de la Distribución Uniforme



Elaborado por: G. Cuenca

La media y la varianza de la distribución uniforme están dadas por:

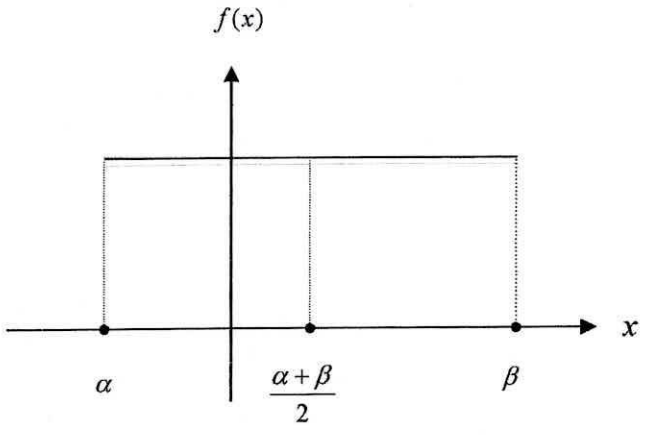
$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{\alpha}^{\beta} x\left(\frac{1}{\beta-\alpha}\right)dx = \left(\frac{x^2}{2(\beta-\alpha)}\right)_{\alpha}^{\beta} = \frac{\beta^2}{2(\beta-\alpha)} - \frac{\alpha^2}{2(\beta-\alpha)}$$

$$= \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{(\beta+\alpha)(\beta-\alpha)}{2(\beta-\alpha)} = \frac{\beta+\alpha}{2} = \mu \qquad (2.2)$$

GRÁFICO 2.2

Efectos de la Imputación en análisis de datos multivariados

Media de la distribución uniforme



Elaborado por: G. Cuenca

Por lo tanto, μ se ubica en el punto medio entre α y β, véase Gráfico 2.2

Además si $X \sim U(\alpha, \beta)$,

$$Var(X) = E(X-\mu)^2 = E(X^2) - \mu^2$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{-\infty}^{\infty} x^2 \left(\frac{1}{\beta-\alpha}\right)dx = \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} x^2 dx = \frac{1}{\beta-\alpha}\left(\frac{x^3}{3}\right)_{\alpha}^{\beta}$$

$$= \frac{1}{\beta-\alpha}\left(\frac{\beta^3}{3} - \frac{\alpha^3}{3}\right) = \frac{1}{\beta-\alpha}\left(\frac{\beta^3-\alpha^3}{3}\right) = \frac{(\beta-\alpha)(\beta^2+\alpha\beta+\alpha^2)}{3(\beta-\alpha)}$$

$$Var(X) = \frac{\beta^2+\alpha\beta+\alpha^2}{3} - \left(\frac{\alpha+\beta}{2}\right)^2 = \frac{\beta^2+\alpha\beta+\alpha^2}{3} - \frac{\alpha^2+2\alpha\beta+\beta^2}{4}$$

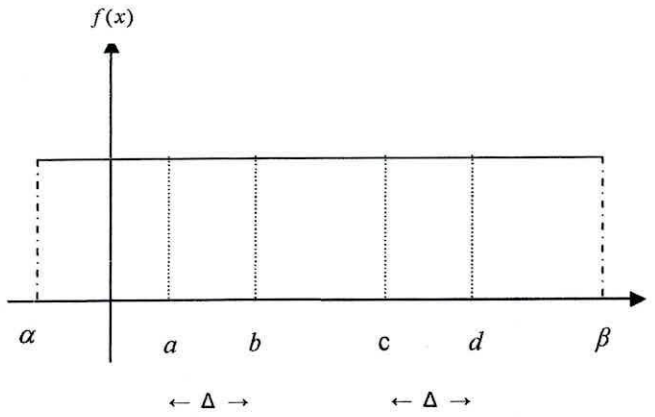
$$= 4(\beta^2+\alpha\beta+\alpha^2) - 3(\alpha^2+2\alpha\beta+\beta^2)$$

$$= \frac{4\beta^2+4\alpha\beta+4\alpha^2-3\alpha^2-6\alpha\beta-3\beta^2}{12}$$

$$= \frac{\beta^2-2\alpha\beta-\alpha^2}{12} = \frac{(\beta-\alpha)^2}{12} \qquad (2.3)$$

Para comprobar que la probabilidad de que un valor ocurra en un intervalo, solo depende de la longitud de dicho intervalo, efectuamos lo siguiente:

GRÁFICO 2.3
Efectos de la Imputación en análisis de datos multivariados
Números en el intervalo $X \in (\alpha, \beta)$



Elaborado por: G. Cuenca

Sea $X \sim U(\alpha, \beta)$, tal que:

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha}, & X \in (\alpha, \beta) \\ 0 \quad , & \textit{para el resto de } X \end{cases}$$

Además, el intervalo (a, b) está incluido en (α, β) al igual que (c, d), esto es:

$$Si\ (a,b) \subseteq (\alpha, \beta) \quad y \quad (c,d) \subseteq (\alpha, \beta)$$

Supongamos además que:

$(b-a) = \Delta x$ y que igualmente $(c-d) = \Delta x$, esto significa que:

$$\mathrm{P}(X \in (a,b)) = \int_a^b \frac{1}{\beta - \alpha} dx = \frac{1}{\beta - \alpha}(b-a) = \frac{\Delta x}{\beta - \alpha}$$

$$\mathrm{P}(X \in (c,d)) = \int_c^d \frac{1}{\beta - \alpha} dx = \frac{1}{\beta - \alpha}(d-c) = \frac{\Delta x}{\beta - \alpha}$$

Por tanto $P(X \in (a,b)) = P(X \in (c,d)) = \dfrac{\Delta x}{\beta - \alpha}$ (2.4)

Caso particular: Si $\alpha = 1$ y $\beta = 10$; $X \sim U(1,10)$

En este caso la densidad f es tal que:

$$f(x) = \begin{cases} \dfrac{1}{9} & X \in (1,10) \\ 0 & \text{resto de } X \end{cases}$$

si $a = 2$, b = 5

$$P(X \in (2,5)) = \int_2^5 \frac{1}{9} dx = \frac{1}{9}(5-2) = \frac{3}{9}$$

si c = 1, d = 4

$$P(X \in (1,4)) = \int_1^4 \frac{1}{9} dx = \frac{1}{9}(4-1) = \frac{3}{9} \quad \text{entonces } \Delta x = 3$$

En este caso particular se puede comprobar que escogiendo números que se encuentran en intervalos de igual longitud $\Delta x = 3$, la probabilidad de que se efectúe una lectura en dicho intervalo es la misma, cuando la variable aleatoria es uniforme.

En cambio si escogemos intervalos de longitud diferente a tres, la probabilidad de que algo ocurra en dicho intervalo, no es $\frac{3}{9}$.

si $h = 1, i = 6$ entonces $\Delta x = 6 - 1 = 5$

$$P(X \in (1,6)) = \int_1^6 \frac{1}{9} dx = \frac{1}{9}(6-1) = \frac{5}{9}$$

La Distribución Acumulada de $X \sim U(\alpha, \beta)$ está dada por:

$$P(X \leq x) = F(x) = \begin{cases} 0 & si \quad X \leq \alpha \\ \dfrac{x-\alpha}{\beta-\alpha} & si \quad \alpha < X < \beta \\ 1 & si \quad X \geq \beta \end{cases} \tag{2.5}$$

Si $X \sim U(0,1)$; $F(x) = \begin{cases} 0 & \text{si} \quad X \leq 0 \\ x & \text{si} \quad X \in (0,1) \\ 1 & \text{si} \quad X \geq 1 \end{cases}$

esto es $F(x) = x$, $x \in (0,1)$ lo cual significa que:

$F(0.10) = P(X \leq 0.10) = 0.10$
$F(0.15) = P(X \leq 0.15) = 0.15$
$F(0.99) = P(X \leq 0.99) = 0.99, etc.$

Por definición la función generadora de momentos de una variable aleatoria continua X es:

$$M_X(t) = E\left[e^{Xt}\right] = \int_{-\infty}^{\infty} e^{Xt} f(x)dx \tag{2.6}$$

la variable independiente es t, y por lo general estamos interesados en los valores de t en una vecindad de cero, por ejemplo $|t| < h$

Ahora se calcula la Función Generadora de Momentos de la Distribución Uniforme

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tX} f(x)dx ,$$

$$M_X(t) = \int_{\alpha}^{\beta} e^{tX}\left(\frac{1}{\beta-\alpha}\right)dx = \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} e^{tX}dx$$

$$= \frac{1}{\beta-\alpha} * \frac{I}{t}\left(e^{tX}\right)_{\alpha}^{\beta} = \frac{1}{\beta-\alpha} * \frac{1}{\mathrm{t}}\left(e^{t\beta} - e^{t\alpha}\right)$$

$$= \frac{e^{t\beta} - e^{t\alpha}}{t(\beta-\alpha)}, \quad t \neq 0 \qquad (2.7)$$

Esto es, $M_X(t)$ no está definida en t=0.

2.3 Prueba de Bondad de Ajuste Ji Cuadrada χ^2

La prueba de bondad de ajuste χ^2 se aplica a situaciones en las que queremos determinar si un conjunto de datos se puede considerar como una muestra aleatoria de una población que tiene una distribución dada. El contraste de hipótesis y el estadístico de prueba utilizados para éste análisis, se presentan en el Cuadro 2.1

Cuadro 2.1

Efectos de la Imputación en el Análisis de Datos Multivariados

Contraste de Hipótesis de la Prueba de Bondad de Ajuste

H_0: La distribución de la población donde se obtuvo la muestra es $F_0(x)$

vs.

H_1: No es verdad H_0

Estadístico de Prueba es : $\sum_{i=1}^{m} \frac{(f_i - e_i)^2}{e_i}$

que sigue una distribución χ^2 y con (m-p-1) grados de libertad

Elaborado por: G. Cuenca

Donde m es el número de términos en la suma y p es el número de parámetros que se estiman en el modelo con base en los datos muestrales.

Con el propósito de ilustrar esta prueba de hipótesis, considere si los siguientes números aleatorios provienen de una distribución Poisson con parámetro conocido $\lambda = 3$

Tabla 2.1
Efectos de la Imputación en el análisis de datos multivariados
Prueba de Bondad de Ajuste

i	Frecuencia Observada f_i	Probabilidad Poisson $\lambda = 3$	e_i
0	18	0.050	22.000
1	53	0.149	65.700
2	103	0.224	98.600
3	107	0.224	98.600
4	82	0.168	73.900
5	46	0.101	44.440
6	18	0.050	22.200
7	10	0.022	9.680
8	3	0.012	5.280

Elaborado por: G. Cuenca

Cuadro 2.2
Efectos de la Imputación en el análisis de datos multivariados
Prueba de Bondad de Ajuste

H_0: La distribución de la población donde se obtuvo la muestra es Poisson $\lambda = 3$

vs.

H_1: No es verdad H_0

Estadístico de Prueba es : $\sum_{i=1}^{m} \frac{(f_i - e_i)^2}{e_i} = 6.828$

Valor p = 0,998

Elaborado por: G. Cuenca

De acuerdo a los resultados obtenidos mediante la prueba de bondad de ajuste, el valor p es 0.998, por lo tanto no existe evidencia estadística suficiente para rechazar la hipótesis nula, es decir, la distribución de la población donde se obtuvo la muestra es Poisson $\lambda = 3$.

2.4 Prueba de Kolmogorov-Smirnov

La prueba de bondad de ajuste KS es una alternativa a la prueba χ^2 que permite comprobar si una muestra aleatoria proviene de una población con una distribución dada, pero se prefiere el uso de la prueba KS en el caso de distribuciones continuas ya que esta prueba trabaja directamente sobre las observaciones y en cambio la prueba χ^2 trabaja sobre los datos agrupados.

Recordemos que dada una muestra aleatoria $X_1, X_2, ..., X_n$ y la muestra ordenada $X_{(1)}, X_{(2)}, ..., X_{(n)}$, la distribución empírica de la muestra es:

$$\hat{F}_n(x) = \begin{cases} 0 & si \quad X < X_{(1)} \\ \dfrac{k}{n} & si \quad X_{(k)} \leq X < X_{(k+1)}; \; si \; k = 1,2,...,n-1 \\ 1 & si \quad X \geq X_{(n)} \end{cases}$$

La prueba KS consiste en verificar el contraste de hipótesis:

Cuadro 2.3
Efectos de la Imputación en el Análisis de Datos Multivariados
Contraste de Hipótesis de la Prueba de Kolmogorov-Smirnov

H_0: La distribución de la población donde se obtuvo la muestra es $F_0(x)$

vs.

H_1: No es verdad $F_0(x)$

Estadístico de Prueba es : $max\left|\hat{F}_n(x) - F_0(x)\right|$

que sigue una distribución D y con (n, p) grados de libertad

Elaborado por: G. Cuenca

Con el propósito de ilustrar esta prueba de hipótesis, se tiene una matriz de datos cuyas columnas son muestras tomadas de cinco poblaciones todas ellas Normal, independientes e idénticamente distribuidas, con parámetros μ=0 y σ^2=1, $\mathbf{X} \in \mathrm{M}_{4x5}$, *i= 1,2,3* y *j= 1,2,3,4,5*

Tabla 2.2
Efectos de la Imputación en el análisis de datos multivariados
Matriz de Datos de variables aleatorias independientes con distribución Normal (0, 1)
Tamaño de muestra n=4

0.464	0.137	2.455	-0.323	-0.068
0.906	-0.513	-0.525	0.595	0.881
-0.482	1.678	-0.057	-1.229	-0.486
-1.787	-0.261	1.237	1.046	-0.508

Elaborado por: G. Cuenca

Tabla 2.3

Efectos de la Imputación en el análisis de datos multivariados

Prueba de Kolmogorov-Smirnov

X_n	$\hat{F}_{20}(x)$	$F_0(x)$	$max\left\|\hat{F}_n(x) - F_0(x)\right\|$
-1.787	1/20	0.037	0.013
-1.229	2/20	0.109	0.009
-0.525	3/20	0.299	0.149
-0.513	4/20	0.305	0.105
-0.508	5/20	0.305	0.056
-0.486	6/20	0.312	0.012
-0.482	7/20	0.316	0.034
-0.323	8/20	0.375	0.026
-0.261	9/20	0.397	0.053
-0.068	10/20	0.472	0.028
-0.057	11/20	0.476	0.074
0.137	12/20	0.556	0.044
0.464	13/20	0.677	0.027
0.595	14/20	0.726	0.026
0.881	15/20	0.811	0.061
0.906	16/20	0.819	0.014
1.046	17/20	0.853	0.003
1.237	18/20	0.893	0.008
1.678	19/20	0.954	0.004
2.455	20/20	0.993	0.006

Elaborado por: G. Cuenca

Cuadro 2.4

Efectos de la Imputación en el análisis de datos multivariados

Prueba de Kolmogorov-Smirnov

H_0: La distribución de la población donde se obtuvo la muestra es Normal(0,1)

vs.

H_1: No es verdad H_0

Estadístico de Prueba es : $max\left|\hat{F}_n(x) - F_0(x)\right|$ = 0.149

Valor p = 0.766

Elaborado por: G. Cuenca

De acuerdo a los resultados obtenidos mediante la prueba de kolmogorov-Smirnov, el valor p es 0.766, por lo tanto no existe evidencia estadística suficiente para rechazar la hipótesis nula, es decir, la distribución de la población donde se obtuvo la muestra es Normal (0, 1).

2.5 Generación de Números Pseudo Aleatorios $U(0,1)$

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. [2]

2.5.1 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.* [2]

2.5.1.1 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:

$$X_{n+1} = (aX_n + c) \bmod m, \tag{2.8}$$

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. [2]

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 2.4 el "período del generador" es ocho, esto es la sucesión se repite una vez que se obtuvo el octavo número generado

TABLA 2.4

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

Elaborado por: G. Cuenca

El procedimiento para obtener los números pseudo aleatorios se realiza de la siguiente forma, donde la semilla es 4:

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

si $n = 0$

$$X_1 = (5X_0 + 7) \bmod 8$$
$$= \frac{5(4)+7}{8} = \frac{27}{8} = 3 + \frac{3}{8} = 3.375$$

donde $\frac{3}{8}$ es el residuo y al dividir 3 para 8, el resultado es el número uniforme 0.375.

si $n = 1$

$$X_2 = (5X_1 + 7) \bmod 8$$
$$= \frac{5(3)+7}{8} = \frac{22}{8} = 2 + \frac{6}{8} = 2.75$$

donde el residuo es $\frac{6}{8}$ y al dividir 6 para 8, el resultado es el número uniforme 0.750.

si $n = 2$

$$X_3 = (5X_2 + 7) \bmod 8$$
$$= \frac{5(6)+7}{8} = \frac{37}{8} = 4 + \frac{5}{8} = 4.62$$

donde el residuo es $\frac{5}{8}$ y al dividir 5 para 8, el resultado es el número uniforme 0.625

y así sucesivamente se calculan los restantes cinco números uniformes de la sucesión, los cuales son: 0.000, 0.875, 0.250, 0.125 y 0.500.

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 2.5. 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 2.5

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

Elaborado por: G. Cuenca

De la misma forma que el ejemplo anterior los números pseudo aleatorios se obtienen de la siguiente manera, donde la semilla es 7:

$$X_{n+1} = (7X_n + 7) \bmod 10$$

si $n = 0$

$$X_1 = (7X_0 + 7) \bmod 10$$
$$= \frac{7(7)+7}{10} = \frac{56}{10} = 5 + \frac{6}{10} = 5.600$$

donde $\frac{6}{10}$ es el residuo y al dividir 6 para 10, el resultado es el número uniforme 0.600

si $n = 1$

$$X_2 = (7X_1 + 7) \bmod 10$$
$$= \frac{7(6)+7}{10} = \frac{49}{10} = 4 + \frac{9}{10} = 4.900$$

donde $\frac{9}{10}$ es el residuo y al dividir 9 para 10, el resultado es el número uniforme 0.900.

si $n = 2$

$$X_3 = (7X_2 + 7) \bmod 10$$
$$= \frac{7(9)+7}{10} = \frac{70}{10} = 7 + \frac{0}{10} = 7.000$$

donde $\frac{0}{10}$ es el residuo y al dividir 0 para 10, el resultado es el número uniforme 0.000

si $n = 3$

$$X_4 = (7X_3 + 7) \bmod 10$$
$$= \frac{7(0)+7}{10} = \frac{7}{10} = 0 + \frac{7}{10} = 0.700$$

donde $\frac{7}{10}$ es el residuo y al dividir 7 para 10, el resultado es el número uniforme 0.700.

Azarang y García expresan:

"Se advierte la necesidad de establecer algunas reglas que pueden ser utilizadas en la selección de los valores de los parámetros, para que el generador resultante tenga período completo". [1]

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.

2.5.1.2 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:

$$X_{n+1} = aX_n \bmod m, \tag{2.9}$$

Al igual que el generador anterior, en éste también se debe seleccionar adecuadamente los valores de los parámetros a, X_0 y m, con el fin de asegurar un período largo para las sucesiones generadas por este método.

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 2.6.

TABLA 2.6

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

Elaborado por: G. Cuenca

Se puede apreciar en Tabla 2.6 que el período del generador es ocho, esto es la sucesión se repite una vez que se obtuvo el octavo número generado.

El procedimiento para obtener los números pseudo aleatorios se realiza de la siguiente forma, donde la semilla es 5.

$$X_{n+1} = 5X_n \bmod 32$$

si $n = 0$

$$X_1 = 5X_0 \bmod 32$$

$$= \frac{5(5)}{32} = \frac{25}{32} = 0 + \frac{25}{32} = 0.781$$

donde $\frac{25}{32}$ es el residuo y al dividir 25 para 32, el resultado es el número uniforme 0.781.

si $n=1$

$$X_2 = 5X_1 \bmod 32$$
$$= \frac{5(25)}{32} = \frac{125}{32} = 3 + \frac{29}{32} = 3.906$$

donde el residuo es $\frac{29}{32}$ y al dividir 29 para 32, el resultado es el número uniforme 0.906.

si $n=2$

$$X_3 = 5X_2 \bmod 32$$
$$= \frac{5(29)}{32} = \frac{145}{32} = 4 + \frac{17}{32} = 4.531$$

donde el residuo es $\frac{17}{32}$ y al dividir 17 para 32, el resultado es el número uniforme 0.531.

y así sucesivamente se calculan los restantes cinco números de la uniformes de la sucesión, los cuales son: 0.656, 0.281, 0.406, 0.031 y 0.156.

2.6 Métodos de Generación de Variables Aleatorias No Uniformes

Generalmente en las simulaciones de sistemas estocásticos existen una o varias variables aleatorias interactuando, las cuales siguen

distribuciones diferentes a la distribución uniforme. Por consiguiente, para simular este tipo de variables, es necesario contar con un generador de números uniformes y una función que transforme estos números uniformes en valores de la distribución de probabilidad deseada. Para esto, se suele utilizar el método de la transformada inversa. [1]

2.6.1 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:

$$\begin{aligned} F(x) &= y \\ F^{-1}(F(x)) &= F^{-1}(y) \\ x &= F^{-1}(y) \end{aligned} \qquad (2.10)$$

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.

2.6.1.1 Distribución exponencial

Utilizando el "Método de la Transformada Inversa" a continuación se desarrolla un generador de variables aleatorias con distribución exponencial. La función de densidad de probabilidad de una variable aleatoria exponencial con parámetro β es:

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases} \tag{2.11}$$

Su función acumulada F es:

$$P(X \leq x) = F(x) = \begin{cases} 1 - e^{-x/\beta} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases} \tag{2.12}$$

Aplicando el método de la transformada inversa, se tiene que si X es exponencial con parámetro β y y uniforme con parámetro 0 y 1:

$$1 - e^{-x/\beta} = y$$
$$e^{-x/\beta} = 1 - y$$

$$\ln e^{-x/\beta} = \ln(1-y)$$

$$x = -\beta \ln(1-y) \tag{2.13}$$

Para ilustrar este generador de variables aleatorias con distribución exponencial, sea $X \sim E(1,\beta)$, tal que $\beta = 2$ y $Y \sim U(0,1)$, si por ejemplo y=0.25.

$$x = -2\ln(1-0.25)$$
$$x = 0.575$$

Este es un valor tomado de una población exponencial.

2.6.2 Procedimientos Especiales

Existen algunas distribuciones como las distribuciones normal, binomial, poisson, etc. cuya simulación a través del método de transformada inversa resulta complicada. Para estas distribuciones es posible utilizar algunas de sus propiedades para facilitar y agilizar el proceso de generación de números aleatorios.

2.6.2.1 Variables que siguen una Distribución Normal

Puesto que no es posible expresar la distribución acumulada de la distribución normal en forma explícita, por ende no se puede utilizar el método de la transformada inversa.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \infty < x < \infty \tag{2.14}$$

Entonces si se desea generar números aleatorios que sigan una Distribución Normal con parámetros conocidos μ y σ^2, se puede hacer uso del Teorema del Límite Central el cual establece que la suma de n variables aleatorias independientes se aproxima a una distribución Normal con media μ y varianza σ^2, a medida que n se aproxima al infinito.

Si por ejemplo, $X_1, X_2, \ldots, X_n$ es una sucesión de n variables aleatorias independientes, tal que $X_i \sim U(0, 1)$, i=1,2,...n

Para la distribución uniforme con parámetros $\alpha = 0$ y $\beta = 1$ se conoce que:

$$\mu = \frac{1}{2} \quad \text{y} \quad \sigma^2 = \frac{1}{12}$$

Aplicando el Teorema del Límite Central a los X_i tenemos que:

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n/12}} \qquad (2.15)$$

Donde Z es un número aleatorio tomado de una población que sigue una distribución Normal con parámetros $\mu = 0$ y $\sigma^2 = 1$.

Para $n = 12$ se obtienen buenas aproximaciones, entonces se tiene que:

$$Z = \frac{\sum_{i=1}^{12} X_i - 12\mu}{\sqrt{12}.\frac{1}{\sqrt{12}}} = \frac{\sum_{i=1}^{12} X_i - 12(1/2)}{\sqrt{12}.\frac{1}{\sqrt{12}}} = \left(\sum_{i=1}^{12} X_i\right) - 6 \qquad (2.16)$$

Recuérdese que $X_i \sim U(0, 1)$

Si se desea obtener números aleatorios X que sigan una distribución Normal con media μ y varianza σ^2, se parte del siguiente resultado, que se relaciona con la distribución normal estándar esto es, $X \sim N(\mu, \sigma^2)$, entonces $Z = \frac{X - \mu}{\sigma}$ es normal estándar.

2.6.2.2 Variables que siguen una Distribución Poisson

Una variable aleatoria X tiene una distribución poisson si y sólo si su distribución de probabilidades está dada por:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{para } x=0, 1, 2, \ldots \qquad (2.17)$$

Entonces la generación de números al azar que sigan una distribución poisson, se lo puede hacer aplicando el método de la transformada inversa.

$$p_{i+1} = P(X = i) = \frac{\lambda}{i+1} p_i; \quad i \geq 0 \qquad (2.18)$$

CAPÍTULO III

3. TÉCNICAS DE IMPUTACIÓN APLICABLES

3.1 Introducción

El propósito del presente capítulo es el de ilustrar las técnicas de imputación para el manejo de datos incompletos en una matriz de datos, para lo cual, en la sección 3.2 se define lo que es "Imputación de datos", la siguiente sección muestra los métodos de "imputación", entre los cuales están, imputación por la media muestral e imputación por regresión.

3.2 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".

3.3 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*.

3.3.1 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:

$$X_{(imp)j} = \frac{\sum_{i=1}^{r} X_{i,(obs)j}}{r} \tag{3.1}$$

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. [6]

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:

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

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

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

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

Primer 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 3.1). Nótese que, un dato faltante representa, en este caso, el 3% de datos faltantes en la matriz de datos.

Tabla 3.1

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

Elaborado por: G. Cuenca

El valor de la media aritmética de X_4, con el dato faltante es $\overline{X}_4 = \frac{6+6+10+2+9+12+9+6+7}{9} = 7.444$, entonces reemplazamos en $X_{74} = 7.444$, así calculamos nuevamente la media aritmética y la varianza con el dato imputado (Ver Cuadro 3.1). El vector de medias de los datos originales es:

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

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

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

Podemos apreciar en el Cuadro 3.1 que la mediana de los datos imputados para X_4 difiere de la mediana de los datos originales o reales, así como de los incompletos, debido a que al incluir la media en los datos incompletos para realizar la estimación, ésta se ubicó en el centro de los datos al ordenarlos junto con el valor de la mediana de los datos incompletos, y como la cantidad de datos es par se procedió a calcular el promedio de los valores antes mencionados donde se obtuvo que el valor de la mediana de X_4 es 7.222.

CUADRO 3.1

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson

Método de Imputación por la Media

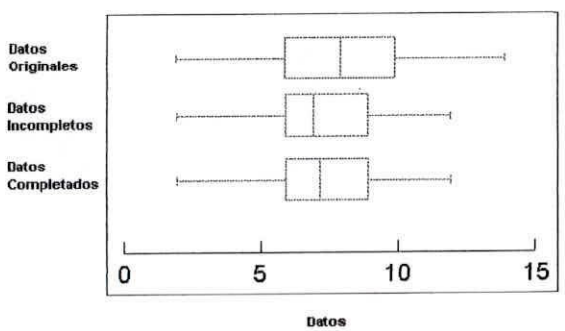
Tamaño de muestra n=10 y 3% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_4"

Estimadores

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media
n		10	9	10
Media		8,100	7,444	7,444
Mediana		8,000	7,000	7,222
Moda		6.000	6,000	6,000
Varianza		11,878	8,528	7,580
Desviación Estándar		3,446	2,920	2,753
Error Estándar		1,090	0,973	0,871
Coeficiente de Asimetría		0,057	-0,334	-0,344
Curtosis		0,150	0,500	0,890
Rango		12,000	10,000	10,000
Mínimo		2,000	2,000	2,000
Máximo		14,000	12,000	12,000
Percentiles	25	6,000	6,000	6,000
	50	8,000	7,000	7,222
	75	10,500	9,500	9,250

Diagrama de Cajas



Elaborado por: G. Cuenca

En el Diagrama de cajas se observa que la distribución de los "datos incompletos", así como de los "datos completados" están sesgadas a la derecha. Para los "datos originales", "incompletos" y "completados", el coeficiente de curtosis es menor a tres, entonces los datos tienen una distribución platicùrtica.

Se puede apreciar también que el valor de la media aritmética de la variable X_4, ($\overline{X}_4 = 7.444$) de los "datos incompletos" y "completados" es igual debido a que si obtenemos el promedio del grupo de datos incompletos y lo agregamos en ese grupo se va obtener el mismo valor del promedio anterior al momento de calcularlo nuevamente. Solo que antes se tenía (n-1) datos y luego n. Pasamos a demostrar esta afirmación:

La media de X_4 con un valor completado es igual a:

$$\overline{X}_{imp} = \frac{X_1 + X_2 + \ldots + X_{n-1} + \dfrac{X_1 + X_2 + \ldots + X_{n-1}}{n-1}}{n}$$

$$= \frac{\dfrac{(n-1)\sum_{i=1}^{n-1} X_i + \sum_{i=1}^{n-1} X_i}{(n-1)}}{n} = \frac{\dfrac{(n-1)\sum_{i=1}^{n-1} X_i + 1\sum_{i=1}^{n-1} X_i}{(n-1)}}{n}$$

$$= \frac{\dfrac{[(n-1)+1]\sum_{i=1}^{n-1} X_i}{(n-1)}}{n} = \frac{\dfrac{n\sum_{i=1}^{n-1} X_i}{(n-1)}}{n} = \frac{\sum_{i=1}^{n-1} X_i}{n-1} = \overline{X}_{n-1}$$

La media para los "datos incompletos" también es igual: $\frac{\sum_{i=1}^{n-1} X_i}{n-1} = \overline{X}_{n-1}$

Queda demostrado que la media para los datos incompletos y de los que tienen como valor imputado la media aritmética de los datos completados, siempre van a ser iguales.

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 3.2).

CUADRO 3.2

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	0.025	0.321	3.543	**7.580**

Elaborado por: G. Cuenca

Por medio del Cuadro 3.2 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. Por otro lado, el valor de las covarianzas disminuyó, con excepción de la covarianza entre X_3 y X_4 donde su valor aumentó de 3.033 a 3.543.

Segundo Caso: Faltan *dos datos* en una misma variable

Como segunda ilustración, utilizamos la misma matriz de datos del primer caso, pero ahora faltan dos datos en la variable X_1, datos que provienen de una distribución Poisson con $\lambda = 2$; faltan: $X_{51} = 4$ y $X_{71} = 2$. Nótese que, dos datos faltantes representan, en este caso, el 5% de datos faltantes en la matriz de datos. (Ver Tabla 3.2)

Tabla 3.2

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, 5% 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

Elaborado por: G. Cuenca

El valor de la media aritmética de X_1, con los dos datos faltantes es $\overline{X}_1 = \frac{5+1+2+2+3+0+3+2}{8} = 2.250$, entonces reemplazamos en $X_{51} = X_{71} = 2.250$.

El vector de medias con dos datos completados en X_1 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \end{pmatrix} = \begin{pmatrix} 2.250 \\ 4.600 \\ 4.700 \\ 8.100 \end{pmatrix}$$

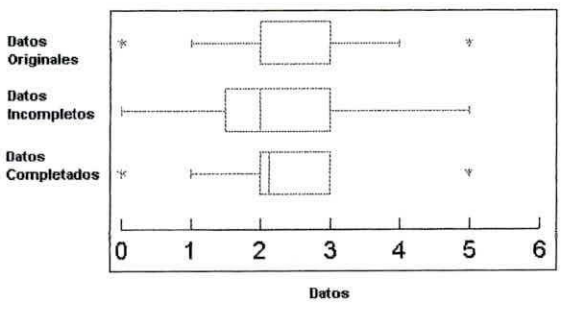
CUADRO 3.3

Efectos de la Imputación en el Análisis de Datos Multivariados
Variables aleatorias independientes con distribución Poisson
Método de Imputación por la Media
Tamaño de muestra n=10 y 5% de datos faltantes en la matriz
Tabla y Diagrama de la "*Variable* X_1"

Estimadores

Estimadores		Datos Originales	Datos Incompletos	Datos completados por la Media
n		10	8	10
Media		2,400	2,250	2,250
Mediana		2,000	2,000	2,125
Moda		2,000	2,000	2,000
Varianza		2,044	2,214	1,728
Desviación Estándar		1,430	1,488	1,312
Error Estándar		0,452	0,526	0,415
Coeficiente de Asimetría		0,251	0,477	0,507
Curtosis		0,341	1,107	1,982
Rango		5,000	5,000	5,000
Mínimo		0,000	0,000	0,000
Máximo		5,000	5,000	5,000
Percentiles	25	1,750	1,250	1,750
	50	2,000	2,000	2,125
	75	3,250	3,000	3,000

Diagrama de Cajas



Elaborado por: G. Cuenca

Podemos apreciar en Cuadro 3.3 que la mediana de los datos con imputación en la variable X_1 es mayor que la de los datos completos e incompletos. Por medio del Diagrama de Cajas se aprecia que las distribuciones de los datos incompletos y con imputación están sesgadas a la derecha, ya que su coeficiente de asimetría es mayor a cero, así como también tienen una distribución leptocùrtica.

La columna con datos originales y con datos imputados tiene valores atípicos estos son 0 y 5.

Analicemos ahora el efecto de esta imputación en la matriz de varianzas y covarianzas, comparando la matriz original con la matriz con 5% de datos completados mediante imputación por la media (Ver Cuadro 3.4)

CUADRO 3.4

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 5% 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
(Dos datos completados con Imputación en X_1)

Variables	X_1	X_2	X_3	X_4
X_1	**1.728**			
X_2	-0.211	2.044		
X_3	-0.436	-0.356	8.900	
X_4	-0.253	-0.844	3.033	11.878

Elaborado por: G. Cuenca

En la matriz de varianzas y covarianzas de los datos completados "con imputación" *por la media* el valor de las covarianzas de la variable X_1

con las demás variables disminuyó, por ejemplo la covarianza entre X_1 y X_2, disminuyó de 0.067 a -0.211.

Tercer Caso: Faltan *cinco datos*, tres en X_1 y dos en X_3

Continuando con la matriz de datos del primer y segundo caso, pero ahora con cinco datos faltantes en total: tres en la variable X_1, datos que provienen de una distribución Poisson con $\lambda = 2$; faltan: $X_{11} = 5$, $X_{51} = 4$ y $X_{71} = 2$ y dos en la variable X_3, datos que provienen de una distribución Poisson con $\lambda = 5$; faltan: $X_{43} = 3$ y $X_{83} = 5$. Nótese que, cinco datos faltantes representan, en este caso, el 13% de datos faltantes en la matriz de datos. (Ver Tabla 3.3)

Tabla 3.3

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, 13% 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

Elaborado por: G. Cuenca

Los valores de las medias aritméticas $\overline{X}_1$ y $\overline{X}_3$ con los datos faltantes, en este caso siete y ocho respectivamente son: 1.857 y 4.875 entonces reemplazamos en los datos faltantes en su respectiva columna de la matriz de datos.

La matriz de datos resultante con cinco valores completados por imputación por media en las variables X_1 y X_3, se muestra en la Tabla 3.4.

Tabla 3.4

Efectos de la imputación en el análisis de datos multivariados
Matriz de datos de variables aleatorias independientes con distribución Poisson
Método de Imputación por la Media
Tamaño de muestra n=10, 13% de datos completados en la matriz

X_1	X_2	X_3	X_4
1.857	4	3	6
1	7	1	6
2	6	8	10
2	5	**4.875**	2
1.857	6	4	9
3	5	6	12
1.857	3	4	14
0	3	**4.875**	9
3	3	2	6
2	4	11	7

Elaborado por: G. Cuenca

El vector de medias con tres datos con imputación en X_1 y dos en X_3 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \end{pmatrix} = \begin{pmatrix} 1.857 \\ 4.600 \\ 4.875 \\ 8.100 \end{pmatrix}$$

CUADRO 3.5

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson

Método de Imputación por la Media

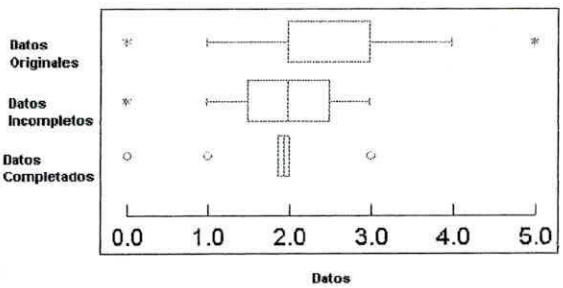
Tamaño de muestra n=10 y 13% de datos faltantes en la matriz

Tablas y Diagramas de las "*Variables* X_1 y X_3"

Estimadores "*Variable* X_1"

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media
n		10	7	10
Media		2,400	1,857	1,857
Mediana		2,000	2,000	1,928
Moda		2,000	2,000	1,857
Varianza		2,044	1,143	0,762
Desviación Estándar		1,430	1,069	0,873
Error Estándar		0,452	0,404	0,276
Coeficiente de Asimetría		0,251	-0,772	-0,844
Curtosis		0,341	0,262	1,619
Rango		5,000	3,000	3,000
Mínimo		0,000	0,000	0,000
Máximo		5,000	3,000	3,000
Percentiles	25	1,750	1,000	1,643
	50	2,000	2,000	1,929
	75	3,250	3,000	2,250

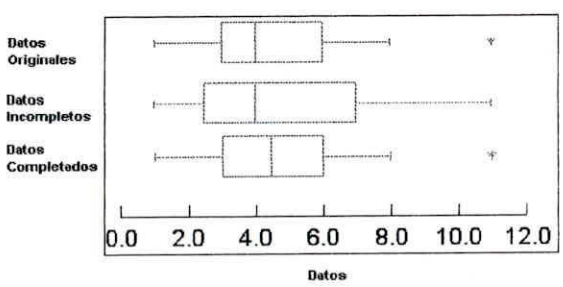
Diagrama de Cajas "*Variable* X_1"



Estimadores "*Variable* X_3"

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media
n		10	8	10
Media		4,700	4,875	4,875
Mediana		4,000	4,000	4,438
Moda		3,000	4,000	4,000
Varianza		8,900	10,982	8,542
Desviación Estándar		2,983	3,314	2,923
Error Estándar		0,943	1,172	0,924
Coeficiente de Asimetría		1,085	0,899	0,956
Curtosis		1,046	0,250	1,080
Rango		10,000	10,000	10,000
Mínimo		1,000	1,000	1,000
Máximo		11,000	11,000	11,000
Percentiles	25	2,750	2,250	2,750
	50	4,000	4,000	4,438
	75	6,500	7,500	6,500

Diagrama de Cajas "*Variable* X_3"



Elaborado por: G. Cuenca

Podemos apreciar en el Cuadro 3.5 que la mediana de los datos con imputación en la primera variable disminuyó ya que existe mayor cantidad de datos con imputación y uno de estos se colocó en el centro al momento de calcular nuevamente la mediana.

En el Diagrama de cajas de la variable X_1 se puede apreciar que tanto los datos originales, incompletos y con imputación tienen un valores atípicos en este caso un valor relativamente pequeño y uno relativamente grande, así como también su distribución es platicùrtica ya que el coeficiente de curtosis de cada una es menor a tres.

El Cuadro 3.5 también nos muestra los estimadores para la variable X_3, donde la mediana de los datos con imputaciòn aumenta con respecto a la mediana de los datos originales e incompletos. El Diagrama de Cajas para los datos originales, incompletos y con imputación muestran que sus distribuciones están sesgadas a la derecha es decir contienen algunos valores relativamente grandes y la curtosis es menor a tres por lo tanto su distribución es platicùrtica.

Analicemos ahora el efecto de esta imputación en la matriz de varianzas y covarianzas, comparando la matriz original con la matriz con 13% de datos completados mediante imputación por la media (Ver Cuadro 3.6)

CUADRO 3.6

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 13% 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 (Datos completados con Imputación en X_1 y X_3)

Variables	X_1	X_2	X_3	X_4
X_1	**0.762**			
X_2	-0.032	2.044		
X_3	0.294	-0.250	**8.542**	
X_4	0.159	-0.844	1.750	11.878

Elaborado por: G. Cuenca

La covarianza entre X_1 y X_4, se incrementa de -0.267 en los datos originales a 0.159 en los datos completados con imputación.

El valor de la covarianza entre X_3 y X_4 disminuye de 3.033 en los datos originales a 1.750 en los datos con imputación.

3.3.2 Modelo de Regresión Lineal Múltiple

Un modelo de Regresión Lineal Múltiple entre una variable dependiente Y y p-1 variables independientes $(X_1, X_2, ..., X_{p-1})$ es un modelo del tipo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i ; \quad \varepsilon_i \sim N(0, \sigma^2) \qquad (3.2)$$

Donde $(X_1, X_2, \ldots, X_{p-1})$ son las variables explicativas de la regresión, Y es la variable explicada y ε es el ruido o error aleatorio.

Los valores de los parámetros β_i son desconocidos y deben ser estimados utilizando una muestra aleatoria que consiste en p-uplas del tipo:

$$\begin{cases} (X_{11}, X_{12}, \ldots, X_{1p}, Y_1) \\ (X_{21}, X_{22}, \ldots, X_{2p}, Y_2) \\ . \\ . \\ . \\ (X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i) \\ . \\ . \\ . \\ (X_{n1}, X_{n2}, \ldots, X_{np}, Y_n) \end{cases}$$

Donde X_{ij} es el valor de la *j*-ésima variable independiente, para la *i*-èsima observación $i = 1,2,..n$

Los resultados se facilitan con el uso de notación matricial con $X_0 = 1$, de la siguiente manera:

$$\mathbf{Y}=\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_i \\ . \\ . \\ . \\ Y_n \end{pmatrix} \quad \mathbf{X}=\begin{pmatrix} 1 & X_{11} & X_{12}....X_{1P} \\ 1 & X_{21} & X_{22}....X_{2P} \\ 1 & X_{31} & X_{32}....X_{3P} \\ . & & \\ . & & \\ 1 & X_{i1} & X_{i2}....X_{iP} \\ . & & \\ . & & \\ 1 & X_{n1} & X_{n2}....X_{nP} \end{pmatrix} \quad \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_i \\ . \\ . \\ \beta_P \end{pmatrix} \quad \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

Por tanto, las n ecuaciones que representan las Y_i como función de las X, los estimadores β y los ε se pueden escribir como:

$$\mathrm{Y}=\mathrm{X}\beta+\varepsilon \tag{3.3}$$

Para n observaciones de un modelo de regresión lineal simple, esto es:

$$Y_i=\beta_0+\beta_1 X_{il}+\varepsilon_i$$

$$\mathbf{X}=\begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ . & \\ . & \\ . & \\ 1 & X_n \end{pmatrix} \quad \mathbf{Y}=\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} \quad \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

Dado que

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} 1 & 1.... & 1 \\ X_1 & X_2 ... X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^{\,2} \end{pmatrix} \tag{3.4}$$

y

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} 1 & 1.... & 1 \\ X_1 & X_2 ... X_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix} \tag{3.5}$$

Vemos que las ecuaciones de Mínimos Cuadrados están dadas por

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{3.6}$$

En consecuencia,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{3.7}$$

Son también las soluciones de Mínimos Cuadrados.

Coeficiente de Correlación Lineal

Una medida para saber que tan adecuado es el modelo lineal general planteado para $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, es el coeficiente de correlación lineal entre X y Y, pero en el cual ya se considera a X como una variable aleatoria.

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3.8}$$

En el modelo de regresión lineal estudiado $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, X es un valor fijado por el investigador, es decir no es variable aleatoria.

$$\text{E(Y)} = \beta_0 + \beta_1 X \tag{3.9}$$

Podemos considerar a x como un valor tomado por cierta variable aleatoria X

$$\text{E}\langle Y \,|\, X = x \rangle = \beta_0 + \beta_1 x \quad \text{donde} \quad \beta_1 = \frac{\sigma_Y}{\sigma_X}\rho$$

Para el caso en que (X, Y) tiene una distribución bivariada, es posible que el investigador no esté interesado en la relación lineal que defina $\text{E}\langle Y \,|\, X \rangle$, éste quizás sólo desee saber si las variables aleatorias X y Y son independientes. Si (X, Y) tiene una distribución normal, entonces la prueba de independencia equivale a probar que el coeficiente de correlación ρ es igual a cero. Se debe recordar que ρ es positivo si X y Y tienden a aumentar juntas y es negativo si ρ disminuye a medida que aumenta X.

3.3.3 Imputación por Regresión

El método de Imputación por Regresión se realiza particionando la matriz $\mathbf{X}$ 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 $\mathbf{X} \in M_{n\times p}$, 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

$$\hat{Y}_j = b_0 + b_1 X_1 + \ldots + b_{j-1} X_{j-1} + b_{j+1} X_{j+1} + \ldots + b_p X_p \qquad (3.10)$$

El cálculo de los coeficientes de la regresión como explicamos en la sección anterior 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. [6]

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, $\mathbf{X} \in \mathrm{M}_{10\mathrm{x}3}$, $i = 1,2,....10$ y $j = 1,2,3$.

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

Tabla 3.5

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

Elaborado por: G. Cuenca

Nótese que, dos datos faltantes representan, en este caso, el 7% de datos faltantes en la matriz de datos. (Ver Tabla 3.5)

Se obtiene la matriz *de varianzas y covarianzas* y *de correlaciones* de la matriz de datos original

CUADRO 3.7

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión

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

Matriz de Varianzas y Covarianzas (Datos Originales)

Variables	X_1	X_2	X_3
X_1	140.509		
X_2	49.759	**72.207**	
X_3	1.948	3.677	**0.250**

Matriz de Correlaciones (Datos Originales)

Variables	X_1	X_2	X_3
X_1	1.000		
X_2	0.494	1.000	
X_3	0.328	**0.865**	1.000

Elaborado por: G. Cuenca

Por medio del Cuadro 3.7 podemos apreciar las varianzas y covarianzas entre las variables, utilizando la matriz de datos originales, donde la mayor covarianza está entre las variables X_1 y X_2, esto es 49.759 seguida por la covarianza entre X_2 y X_3, 3.677.

En la matriz de correlaciones, se nota que la mayor correlación se da entre las variables X_2 y X_3 (0.865), seguida por 0.494 entre las variables X_1 y X_2.

1° Paso

Particionamos la matriz de datos en dos partes:

Tabla 3.6

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

Elaborado por: G. Cuenca

Una parte de la matriz tiene filas con valores completos y la otra parte tiene filas con valores faltantes (Ver Tabla 3.6)

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 $\hat{X}_2 = b_0 + b_1 X_1 + b_3 X_3$;

CUADRO 3.8

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión(Variable Dependiente X_2)

Análisis de Regresión

	Coeficientes	Error Estàndar	t	p
Constante	-39,840	11,742	-3,398	0,019
b_1	0,154	0,171	0,904	0,407
b_3	13,801	4,140	3,328	0,021

Elaborado por: G. Cuenca

Por medio del Cuadro 3.8, podemos ver que los valores de los coeficientes son: $b_0 = -39.840, b_1 = 0.154, b_3 = 13.801$

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. (Ver Tabla 3.7)

Utilizando la ecuación $\hat{X}_2 = b_0 + b_1 X_1 + b_3 X_3$ obtenemos el nuevo valor:

Tabla 3.7

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión

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

Primeros valores estimados

X_1	X_2	X_3
35.011	**4.208**	2.801
35.002	4.901	**3.099**
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

Elaborado por: G. Cuenca

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

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

De igual forma obtenemos la ecuación para X_3 que nos da el nuevo valor de predicción

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

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

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. (Ver Tabla 3.8)

Tabla 3.8

Efectos de la imputación en el análisis de datos multivariados
Matriz de datos de variables aleatorias dependientes con distribución Normal
Método de Imputación por Regresión
Tamaño de muestra n=10, 7% de datos faltantes en la matriz
Segundos valores estimados

X_1	X_2	X_3
35.011	**3.696**	2.801
35.002	4.901	**2.978**
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

Elaborado por: G. Cuenca

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

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

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

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

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 $\overline{X}_3 = 3.134$. (Ver Cuadro 3.9)

CUADRO 3.9

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{1,2}$=3.500

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	4.208	0.708
2	3.696	0.196
3	3.702	0.202

Imputaciones sucesivas para $X_{2,3}$=2.702

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	3.028	0.326
2	2.978	0.276
3	2.981	0.279

Elaborado por: G. Cuenca

La matriz de varianzas y covarianzas para datos originales, datos con primera imputaciòn y datos con segunda imputación se muestra a continuación:

CUADRO 3.10

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión

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

Matriz de Varianzas y Covarianzas

(Datos Originales)

Variables	X_1	X_2	X_3
X_1	140.509		
X_2	49.759	72.207	
X_3	1.948	3.677	0.250

Matriz de Varianzas y Covarianzas

(Datos con primer resultado de predicción en $X_{1,2}$ y $X_{2,3}$)

Variables	X_1	X_2	X_3
X_1	140.509		
X_2	50.300	71.677	
X_3	2.251	3.550	0.232

Matriz de Varianzas y Covarianzas

(Datos con segundo resultado de predicción en $X_{1,2}$ y $X_{2,3}$)

Variables	X_1	X_2	X_3
X_1	140.509		
X_2	49.909	72.050	
X_3	2.159	3.600	0.234

Matriz de Varianzas y Covarianzas

(Datos con tercer resultado de predicción en $X_{1,2}$ y $X_{2,3}$)

Variables	X_1	X_2	X_3
X_1	140.509		
X_2	49.927	72.032	
X_3	2.161	3.598	0.234

Elaborado por: G. Cuenca

En el Cuadro 3.10 se aprecia que, con el primer resultado de predicciòn la covarianza entre X_1 y X_2 se incrementa de 49.759 a 50.300, Así como también la covarianza entre X_1 y X_3 de 1.948 a 2.251.

Mientras que las covarianzas con el segundo resultado de predicción empiezan a disminuir es decir, la covarianza entre X_1 y X_2 diminuye de 50.300 a 49.909.

La covarianza entre X_1 y X_2 con el tercer resultado de predicción, disminuye a 49.927 pero este valor tiende al de la matriz de datos originales.

Segundo Caso: Faltan *tres datos*, uno en X_1, uno en X_2 y uno X_3, pero todos pertenecen a una misma fila.

Como ya lo explicamos anteriormente, cuando se da el caso donde no se tiene información suficiente para calcular la ecuación de predicción inicial, se aplica primero el método de imputación por la media y luego se usa regresión en subsecuentes iteraciones.

Utilizando la matriz del caso 1, es decir 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*, y se supone que tiene el 10% de datos faltantes, es decir tres datos, los que recayeron en la variable X_1, X_2 y X_3 : $X_{2,1}$=35.002, $X_{2,2}$=4.901 y el $X_{2,3}$=2.702 (Ver Tabla 3.9)

Tabla 3.9

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, 10% 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

Elaborado por: G. Cuenca

Como podemos observar en la Tabla 3.9, no se tiene suficiente información para obtener la ecuación de predicción, ya que para obtener la misma se requieren que las otras variables tengan datos completos, entonces se procede primero a aplicar el método de imputación por la media.

Los valores de las medias aritméticas $\overline{X}_1$, $\overline{X}_2$ y $\overline{X}_3$ utilizando solo los datos completos, en este caso nueve para cada una son: 27.371, 7.445 y 3.134 entonces reemplazamos en los datos faltantes de su respectiva variable. (Ver Tabla 3.10)

Tabla 3.10

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Método de Imputación por Media

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

X_1	X_2	X_3
35.011	3.500	2.801
27.371	**7.445**	**3.134**
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

Elaborado por: G. Cuenca

Ya que estimamos los valores faltantes primero por medio del método de imputación por la media, ahora procedemos a aplicar imputación por regresión en los mismos.

Primero X_2 y X_3 son las variables independientes que van a explicar a X_1; utilizando la ecuación de regresión $\hat{X}_1 = b_0 + b_2 X_2 + b_3 X_3$

CUADRO 3.11

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión (Variable dependiente X_1)

Análisis de Regresión

	Coeficientes	Error Estàndar	t	p
Constante	37.377	44.259	0.845	0.426
b_2	1.000	0.906	1.103	0.306
b_3	-5.569	-0.350	-0.350	0.737

Elaborado por: G. Cuenca

Entonces $b_0 = 37.377, b_1 = 1.000, b_3 = -5.569$

$$\hat{X}_1 = (37.377) + (1.000)(7.445) + (5.569)(3.134)$$
$$\hat{X}_1 = 27.371$$

De manera similar hacemos la siguiente regresión pero ahora X_1 y X_3 son las variables independientes que van a explicar a X_2, donde

$b_0 = -40.292, b_1 = 0.148, b_3 = 13.940$

CUADRO 3.12

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión (Variable dependiente X_2)

Análisis de Regresión

	Coeficientes	Error Estàndar	t	p
Constante	-40.292	9.372	-4.299	0.004
b_1	0.148	0.134	1.103	0.306
b_3	13.940	3.237	4.306	0.004

Elaborado por: G. Cuenca

$$\hat{X}_2 = (-40.292) + (0.148)(27.371) + (13.940)(3.134)$$
$$\hat{X}_2 = 7.443$$

Por último hacemos regresión donde, X_1 y X_2 son las variables independientes que van a explicar a X_3

$b_0 = 2.830, b_1 = -0.003, b_3 = -0.052$

$$\hat{X}_3 = (2.830) + (-0.003)(27.371) - (0.052)(7.445)$$
$$\hat{X}_3 = 2.358$$

CUADRO 3.13

Efectos de la Imputación en el Análisis de Datos Multivariados
Variables aleatorias dependientes con distribución Normal
Método de Imputación por Regresión (Variable dependiente X_3)

Análisis de Regresión

	Coeficientes	Error Estàndar	t	p
Constante	2.830	0.224	12.626	0.000
b_1	-0.003	0.009	-0.350	0.737
b_3	-0.052	0.012	4.306	0.004

Elaborado por: G. Cuenca

Ahora insertamos estos estimadores, 27.371 en X_{21}, 7.443 en X_{22} y 2.358 en X_{23}, (Ver Tabla 3.11) y calculamos nuevamente la ecuación de regresión.

Tabla 3.11

Efectos de la imputación en el análisis de datos multivariados
Matriz de datos de variables aleatorias dependientes con distribución Normal
Método de Imputación por Regresiòn
Tamaño de muestra n=10, 10% de datos faltantes en la matriz
Primeros valores estimados

X_1	X_2	X_3
35.011	3.500	2.801
27.371	**7.443**	**2.358**
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

Elaborado por: G. Cuenca

$$\hat{X}_1 = (29.060) + (0.854)(7.442) + (-2.633)(2.358)$$
$$\hat{X}_1 = 29.207$$

$$\hat{X}_2 = (-30.629) + (0.204)(27.371) + (10.630)(2.358)$$
$$\hat{X}_2 = 0.020$$

$$\hat{X}_3 = (2.753) + (-0.003)(27.371) - (0.052)(7.443)$$
$$\hat{X}_3 = 2.281$$

Insertamos los nuevos estimadores, 29.207 en X_{21}, 0.020 en X_{22} y 2.281 en X_{23}, para calcular los nuevos valores de predicción. (Ver Tabla 3.12)

Tabla 3.12

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresiòn

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

Segundos valores estimados

X_1	X_2	X_3
35.011	3.500	2.801
29.207	**0.020**	**2.281**
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

Elaborado por: G. Cuenca

$$\hat{X}_1 = (45.307) + (1.103)(0.020) + (-8.248)(2.281)$$
$$\hat{X}_1 = 26.713$$

$$\hat{X}_2 = (-36.469) + (0.175)(29.207) + (12.578)(2.281)$$
$$\hat{X}_2 = -2.666$$

$$\hat{X}_3 = (2.824) + (-0.006)(29.207) + (0.058)(0.020)$$
$$\hat{X}_3 = 2.647$$

Los nuevos estimadores son insertados, 26.713 en X_{21}, -2.666 en X_{22} y 2.647 en X_{23}, para calcular los nuevos valores de predicción. (Ver Tabla 3.13)

Tabla 3.13

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresiòn

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

Terceros valores estimados

X_1	X_2	X_3
35.011	3.500	2.801
26.713	**-2.666**	**2.647**
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

Elaborado por: G. Cuenca

$$\hat{X}_1 = (38.052) + (0.907)(-2.666) + (-5.375)(2.647)$$
$$\hat{X}_1 = 21.405$$

$$\hat{X}_2 = (-42.522) + (0.138)(26.713) + (14.652)(2.647)$$
$$\hat{X}_2 = -0.047$$

$$\hat{X}_3 = (2.832) + (-0.003)(26.713) + (0.051)(-2.666)$$
$$\hat{X}_3 = 2.687$$

Los nuevos estimadores insertados son, 21.405 en X_{21}, -0.047 en X_{22} y 2.687 en X_{23}, para calcular los nuevos valores de predicción. (Ver Tabla 3.14)

Tabla 3.14

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión

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

Cuartos valores estimados

X_1	X_2	X_3
35.011	3.500	2.801
21.405	**-0.047**	**2.687**
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

Elaborado por: G. Cuenca

$$\hat{X}_1 = (36.889) + (1.003)(-0.047) + (-5.449)(2.687)$$
$$\hat{X}_1 = 22.199$$

$$\hat{X}_2 = (-40.538) + (0.149)(21.405) + (14.002)(2.687)$$
$$\hat{X}_2 = 0.275$$

$$\hat{X}_3 = (2.817) + (-0.003)(21.405) + (0.053)(-0.047)$$
$$\hat{X}_3 = 2.749$$

Los nuevos estimadores que se insertan son, 22.199 en X_{21}, 0.275 en X_{22} y 2.749 en X_{23}, para calcular los nuevos valores de predicción. (Ver Tabla 3.15)

Tabla 3.15

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresiòn

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

Quintos valores estimados

X_1	X_2	X_3
35.011	3.500	2.801
22.199	**0.275**	**2.749**
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

Elaborado por: G. Cuenca

$$\hat{X}_1 = (37.336) + (1.001)(0.275) + (-5.563)(2.749)$$
$$\hat{X}_1 = 22.316$$

$$\hat{X}_2 = (-40.897) + (0.149)(22.199) + (14.091)(2.749)$$
$$\hat{X}_2 = 1.152$$

$$\hat{X}_3 = (2.826) + (-0.003)(22.199) + (0.052)(0.275)$$
$$\hat{X}_3 = 2.772$$

Entonces los estimadores que se insertarán son, 22.316 en X_{21}, 1.152 en X_{22} y 2.772 en X_{23}, para calcular los nuevos valores de predicción. (Ver Tabla 3.16)

Tabla 3.16

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresiòn

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

Sextos valores estimados

X_1	X_2	X_3
35.011	3.500	2.801
22.316	**1.152**	**2.772**
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

Elaborado por: G. Cuenca

$$\hat{X}_1 = (37.075) + (1.003)(1.152) + (-5.504)(2.772)$$
$$\hat{X}_1 = 22.974$$

$$\hat{X}_2 = (-40.570) + (0.149)(22.316) + (14.007)(2.772)$$
$$\hat{X}_2 = 1.583$$

$$\hat{X}_3 = (2.822) + (-0.003)(22.316) + (0.052)(1.152)$$
$$\hat{X}_3 = 2.814$$

Continuamos insertando los estimadores que ahora, 22.974 en X_{21}, 1.583 en X_{22} y 2.814 en X_{23}. (Ver Tabla 3.17)

Tabla 3.17

Efectos de la imputación en el análisis de datos multivariados

Matriz de datos de variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresiòn

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

Séptimos valores estimados

X_1	X_2	X_3
35.011	3.500	2.801
22.974	**1.583**	**2.814**
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

Elaborado por: G. Cuenca

$$\hat{X}_1 = (37.287) + (1.002)(1.583) + (-5.504)(2.814)$$
$$\hat{X}_1 = 23.243$$

$$\hat{X}_2 = (-40.675) + (0.149)(22.974) + (14.032)(2.814)$$
$$\hat{X}_2 = 2.237$$

$$\hat{X}_3 = (2.826) + (-0.003)(22.974) + (0.052)(1.583)$$
$$\hat{X}_3 = 2.838$$

Se continúa con las regresiones sucesivas, la cual se estabilizó en la iteración treinta y uno (Ver Cuadro 3.14, 3.15 y 3.16), es decir se tuvo

que realizar treinta y un regresiones sucesivas hasta que al final quedaron los siguientes valores estimados para cada variable; 29.547 en $\hat{X}_{21}$, 6.382 en $\hat{X}_{22}$ y 2.347 en $\hat{X}_{23}$.

CUADRO 3.14

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{2,1}$=35.002

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	27.371	7,631
2	29.207	5,795
3	26.713	8,289
4	21.405	13,597
5	22.199	12,803
6	22.136	12,866
7	22.974	12,028
8	23.731	11,271
9	24.008	10,994
10	24.630	10,372
11	24.931	10,071
12	25.366	9,636
13	25.731	9,271
14	26.145	8,857
15	26.351	8,651
16	27.105	7,897
17	27.542	7,460
18	28.372	6,630
19	28.758	6,244
20	29.216	5,786
21	29.843	5,159
22	30.280	4,722
23	30.874	4,128
24	31.520	3,482
25	32.341	2,661
26	32.782	2,220
27	33.451	1,551
28	33.894	1,108
29	34.247	0,755
30	34.784	0,218
31	34.985	0,017

Elaborado por: G. Cuenca

La diferencia en valor absoluto entre el dato observado y el último resultado de predicción tiende al dato observado.

CUADRO 3.15

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{2,2}$=4.901

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	6.382	1,481
2	6.352	1,451
3	6.327	1,426
4	6.305	1,404
5	6.237	1,336
6	6.256	1,355
7	6.220	1,319
8	6.201	1,300
9	6.168	1,267
10	6.120	1,219
11	6.005	1,104
12	5.903	1,002
13	5.856	0,955
14	5.824	0,923
15	5.792	0,891
16	5.741	0,840
17	5.703	0,802
18	5.693	0,792
19	5.637	0,736
20	5.502	0,601
21	5.426	0,525
22	5.315	0,414
23	5.226	0,325
24	5.101	0,200
25	5.003	0,102
26	4.982	0,081
27	4.972	0,071
28	4.958	0,057
29	4.935	0,034
30	4.924	0,023
31	4.910	0,009

Elaborado por: G. Cuenca

CUADRO 3.16

Efectos de la Imputación en el Análisis de Datos Multivariados
Variables aleatorias dependientes con distribución Normal
Método de Imputación por Regresión
Tamaño de muestra n=10, 10% de datos faltantes en la matriz

Imputaciones sucesivas para $X_{2,3}$=2.702

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	2.358	0,344
2	2.281	0,421
3	2.647	0,055
4	2.687	0,015
5	2.749	0,047
6	2.772	0,070
7	2.814	0,112
8	2.838	0,136
9	2.870	0,168
10	2.892	0,190
11	2.917	0,215
12	2.936	0,234
13	2.957	0,255
14	2.972	0,270
15	2.989	0,287
16	3.001	0,299
17	3.014	0,312
18	3.026	0,324
19	3.036	0,334
20	3.045	0,343
21	3.054	0,352
22	3.003	0,301
23	3.891	1,189
24	2.805	0,103
25	2.792	0,090
26	2.772	0,070
27	2.754	0,052
28	2.742	0,040
29	2.731	0,029
30	2.711	0,009
31	2.705	0,003

Elaborado por: G. Cuenca

La diferencia en valor absoluto entre el dato observado de cada variable (X_{21} = 35.002, X_{22} = 4.901 y X_{23} = 2.702), y el último

resultado de predicción por medio del método de imputación por regresión ($\hat{X}_{21}$ = 29.547, $\hat{X}_{22}$ = 6.382 y $\hat{X}_{23}$ = 2.347), es el siguiente:

$$\left|X_{21} - \hat{X}_{21}\right| = \left|35.002 - 34.985\right| = 0.017$$

$$\left|X_{22} - \hat{X}_{22}\right| = \left|4.901 - 4.910\right| = 0.009$$

$$\left|X_{23} - \hat{X}_{23}\right| = \left|2.702 - 2.705\right| = 0.003$$

Mientras que el error entre el dato observado de cada variable (X_{21} = 35.002, X_{22} = 4.901 y X_{23} = 2.702), y el dato estimado por medio del método de imputación por la media ($\hat{X}_{21}$ = 27.371, $\hat{X}_{22}$ = 7.445 y $\hat{X}_{23}$ = 3.134), es el siguiente:

$$\left|X_{21} - \hat{X}_{21}\right| = \left|35.002 - 27.371\right| = 7.631$$

$$\left|X_{22} - \hat{X}_{22}\right| = \left|4.901 - 7.445\right| = 2.544$$

$$\left|X_{23} - \hat{X}_{23}\right| = \left|2.702 - 3.134\right| = 0.432$$

Como podemos apreciar, que la diferencia en valor absoluto entre el dato observado y el estimado por medio del método de imputación por regresión es menor al del estimado por el método de imputación por la media, es decir estos valores tienden a los datos observados.

Analicemos el efecto que causa en la matriz de varianzas y covarianzas,

CUADRO 3.17

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal

Método de Imputación por Regresión

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

Matriz de Varianzas y Covarianzas
(Datos Originales)

Variables	X_1	X_2	X_3
X_1	140.509		
X_2	49.759	72.207	
X_3	1.948	3.677	0.250

Matriz de Varianzas y Covarianzas
(Datos Completados por la Media en X_1, X_2 y X_3)

Variables	X_1	X_2	X_3
X_1	134.685		
X_2	51.700	71.560	
X_3	2.277	3.567	0.232

Matriz de Varianzas y Covarianzas
(Datos Completados por Regresión en X_1, X_2 y X_3)

Variables	X_1	X_2	X_3
X_1	135.159		
X_2	51.469	71.673	
X_3	2.063	3.672	0.329

Elaborado por: G. Cuenca

Las varianzas y covarianzas entre las variables (Ver Cuadro 3.17), utilizando la matriz de datos originales, son las siguientes:

Se aprecia un valor grande en la varianza de la variable X_1 (140.509), en la matriz de varianzas y covarianzas de datos originales, entonces los valores de esta variable tienden a distribuirse lejos de la media, mientras que en la variable X_3 se aprecia una varianza pequeña (0.250), es decir los valores de esta variable tienden a distribuirse cerca de la media.

En la matriz de varianzas y covarianzas de los datos completados por la media se nota una disminución en el valor de las varianzas de las variables, comparándola con la matriz de datos original; esto ocurre debido a que se inserta el valor de la media en los datos faltantes y por ende los datos están menos dispersos. Mientras que el valor de las covarianzas aumentó, con excepción de la covarianza entre las variables X_2 y X_3 donde su valor disminuyó de 3.677 a 3.567.

En la matriz de varianzas y covarianzas de los datos completados por la regresión el valor de las varianzas de las variables aumentó, comparándolo con los datos imputados por la media; mientras que el valor de las covarianzas disminuyó, con excepción de la covarianza entre las variables X_2 y X_3 donde su valor aumentó de 3.567 a 3.672.

CAPÍTULO IV

4. SIMULACIÓN BAJO DISTINTAS CONDICIONES UNIVARIADAS Y MULTIVARIADAS

4.1 Introducción

En el presente capítulo se presentan y analizan los resultados obtenidos al comparar los métodos de imputación utilizando diferentes tamaños de muestras: 30, 50 y 100 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. Para la generación de las variables aleatorias se utiliza el programa Matlab 6.5 el cual provee de los comandos adecuados para la realización de esta tarea.

Se escogieron tamaños de muestra de 30, 50 y 100, puesto que en primera instancia se realizó simulaciones con tamaños de muestra *n*=10, de los cuales no se pudo obtener resultados dignos de comentario.

En la sección 4.2 se presentan simulaciones para distribuciones normal, poisson y exponencial, idénticamente distribuìdas e independientes, mientras que en la sección 4.3 se presentan distribuciones con variables aleatorias dependientes. Para la utilización del Método de Imputación por Regresión se desarrolló un algoritmo en Matlab 6.5 (Ver Anexo 2).

4.2 Matrices de Datos con variables aleatorias independientes

4.2.1 Distribución Normal: *Tres datos faltantes* en una sola variable (2% de la matriz), tamaño de muestra n=30

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, $\mathbf{X} \in \mathrm{M}_{30\mathrm{x}5}$, *i= 1,2,....30* y *j= 1,2,3,4,5* y se supone que tiene el 2% 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 2% de datos faltantes en la matriz, constituye 10% de datos faltantes en la columna que corresponde a X_1 (Ver Tabla 4.1).

Tabla 4.1

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Normal (5, 1)

Tamaño de muestra n=30

X_1	X_2	X_3	X_4	X_5
4.813	3.396	5.569	3.812	5.806
5.726	5.257	4.744	2.798	5.232
4.412	3.944	4.623	5.986	4.010
7.183	6.415	4.704	4.481	6.340
4.864	4.195	3.525	5.327	5.290
5.114	5.529	4.766	5.234	6.479
6.067	5.219	5.118	5.022	6.138
5.059	4.078	5.315	3.996	4.316
4.904	2.829	6.444	4.053	3.708
4.168	4.941	4.649	4.626	4.927
5.294	3.989	5.623	3.814	4.669
3.664	5.615	5.799	3.944	4.156
5.714	5.508	5.941	6.473	5.498
6.624	6.692	4.008	5.056	6.489
4.308	5.591	5.212	3.783	4.454
5.858	4.356	5.238	4.959	4.153
6.254	5.380	3.992	3.872	4.754
3.406	3.991	4.258	3.651	5.663
3.559	4.981	6.082	4.739	4.146
5.571	4.952	4.869	5.954	3.799
4.600	5.000	5.390	5.129	4.880
5.690	4.682	5.088	5.657	4.935
5.816	6.095	4.365	3.832	5.485
5.712	3.126	4.440	4.539	4.405
6.290	5.428	5.444	4.738	4.850
5.669	5.896	4.050	3.787	4.565
6.191	5.731	5.781	3.681	4.921
3.798	5.578	5.569	5.931	6.535
4.980	5.040	4.178	5.011	4.394
4.843	5.677	4.734	4.355	3.653

Elaborado por: G. Cuenca

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 Eliminación por Filas

Como detallamos en el Capítulo 1, el Método de Eliminación por Filas no toma en cuenta las filas donde se encuentren datos faltantes y como los datos faltantes 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, se procede a prescindir de las filas diez, catorce y veinte y cinco. La matriz resultante con filas eliminadas se muestra en la Tabla 4.2.

Tabla 4.2

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Normal (5,1)

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Matriz de datos con tres filas eliminadas

X_1	X_2	X_3	X_4	X_5
4.813	3.396	5.569	3.812	5.806
5.726	5.257	4.744	2.798	5.232
4.412	3.944	4.623	5.986	4.010
7.183	6.415	4.704	4.481	6.340
4.864	4.195	3.525	5.327	5.290
5.114	5.529	4.766	5.234	6.479
6.067	5.219	5.118	5.022	6.138
5.059	4.078	5.315	3.996	4.316
4.904	2.829	6.444	4.053	3.708
5.294	3.989	5.623	3.814	4.669
3.664	5.615	5.799	3.944	4.156
5.714	5.508	5.941	6.473	5.498
4.308	5.591	5.212	3.783	4.454
5.858	4.356	5.238	4.959	4.153
6.254	5.380	3.992	3.872	4.754
3.406	3.991	4.258	3.651	5.663
3.559	4.981	6.082	4.739	4.146
5.571	4.952	4.869	5.954	3.799
4.600	5.000	5.390	5.129	4.880
5.690	4.682	5.088	5.657	4.935
5.816	6.095	4.365	3.832	5.485
5.712	3.126	4.440	4.539	4.405
5.669	5.896	4.050	3.787	4.565
6.191	5.731	5.781	3.681	4.921
3.798	5.578	5.569	5.931	6.535
4.980	5.040	4.178	5.011	4.394
4.843	5.677	4.734	4.355	3.653

Elaborado por: G. Cuenca

Nótese que la eliminación por filas, equivale a prescindir de todos los datos de los informantes porque no respondieron, por ejemplo, una pregunta.

El vector de medias para las veintisiete filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.114 \\ 5.040 \\ 5.088 \\ 4.481 \\ 4.754 \end{pmatrix}$$

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

Ahora analicemos en el Cuadro 4.1 el efecto que causa en la *matriz de varianzas y covarianzas*, y *matriz de correlaciones*, la eliminación de tres filas, es decir la diez, la catorce y la veinticinco, con un tamaño de muestra n=30.

Se puede notar que la mayoría de las covarianzas entre las variables tanto en la matriz de datos originales como en la matriz con tres filas eliminadas son cercanas a cero, lo cual era de esperarse dado que las columnas son muestras tomadas de poblaciones independientes.

CUADRO 4.1

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Normal (5,1)

Método de Eliminación por Filas

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	0.891				
X_2	0.299	0.891			
X_3	-0.152	-0.138	0.502		
X_4	-0.010	0.034	0.014	0.756	
X_5	0.197	0.315	-0.123	0.090	0.740

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.335	1.000			
X_3	-0.227	-0.206	1.000		
X_4	-0.012	0.042	0.023	1.000	
X_5	0.242	**0.388**	-0.202	0.120	1.000

Matriz de Varianzas y Covarianzas (3 Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	0.827				
X_2	0.214	0.866			
X_3	-0.147	-0.095	0.510		
X_4	-0.041	0.004	0.031	0.835	
X_5	0.136	**0.247**	-0.077	0.073	0.732

Matriz de Correlaciones (3 Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.253	1.000			
X_3	-0.226	-0.143	1.000		
X_4	-0.050	0.005	0.048	1.000	
X_5	0.175	**0.311**	-0.125	0.094	1.000

Elaborado por: G. Cuenca

La mayor covarianza en la matriz de datos originales se da entre las variables X_2 y X_5 y es 0.315; mientras que en la matriz con tres filas eliminadas este valor disminuye a 0.247.

En la matriz de correlaciones de datos originales, la mayor correlación se da entre las variables X_2 y X_5, y es 0.388, la que disminuye a 0.311 en la matriz de correlaciones con tres filas eliminadas.

En el Cuadro 4.2, podemos apreciar que con el 10% de datos eliminadas en la primera columna (Variable X_1), el valor de la varianza disminuyó de 0.891 a 0.827.

CUADRO 4.2

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Normal (5,1)

Método de Eliminación por Filas

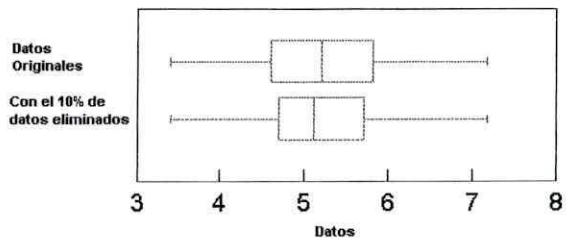
Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_1"

Estimadores

Estimadores		Datos Originales	Con el 10% de datos eliminadas en X_1
n		30	27
Media		5,205	5,151
Mediana		5,205	5,114
Moda		3,410	3,410
Varianza		0,891	0,827
Desviación Estándar		0,944	0,909
Error Estándar		0,172	0,175
Coeficiente de Asimetría		-0,161	-0,176
Curtosis		-0,460	-0,105
Rango		3,780	3,780
Mínimo		3,410	3,410
Máximo		7,180	7,180
Percentiles	25	4,553	4,600
	50	5,204	5,114
	75	5,827	5,726

Diagrama de Cajas



Elaborado por: G. Cuenca

Método de Imputación por la Media y Regresión

A continuación se aplica el método de imputación por media y regresión a la misma matriz de datos originales, con los mismos datos faltantes, utilizada en el método de eliminación por filas.

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 $X_{25,1}$. La matriz de datos resultante con tres valores completados por imputación por la media en la variable X_1 se muestra en la Tabla 4.3.

Tabla 4.3

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Normal (5, 1)

Método de Imputación por la Media

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
4.813	3.396	5.569	3.812	5.806
5.726	5.257	4.744	2.798	5.232
4.412	3.944	4.623	5.986	4.010
7.183	6.415	4.704	4.481	6.340
4.864	4.195	3.525	5.327	5.290
5.114	5.529	4.766	5.234	6.479
6.067	5.219	5.118	5.022	6.138
5.059	4.078	5.315	3.996	4.316
4.904	2.829	6.444	4.053	3.708
5.151	4.941	4.649	4.626	4.927
5.294	3.989	5.623	3.814	4.669
3.664	5.615	5.799	3.944	4.156
5.714	5.508	5.941	6.473	5.498
5.151	6.692	4.008	5.056	6.489
4.308	5.591	5.212	3.783	4.454
5.858	4.356	5.238	4.959	4.153
6.254	5.380	3.992	3.872	4.754
3.406	3.991	4.258	3.651	5.663
3.559	4.981	6.082	4.739	4.146
5.571	4.952	4.869	5.954	3.799
4.600	5.000	5.390	5.129	4.880
5.690	4.682	5.088	5.657	4.935
5.816	6.095	4.365	3.832	5.485
5.712	3.126	4.440	4.539	4.405
5.151	5.428	5.444	4.738	4.850
5.669	5.896	4.050	3.787	4.565
6.191	5.731	5.781	3.681	4.921
3.798	5.578	5.569	5.931	6.535
4.980	5.040	4.178	5.011	4.394
4.843	5.677	4.734	4.355	3.653

Elaborado por: G. Cuenca

Por medio del Método de *Imputación por Regresión*, el cálculo de los valores faltantes se realiza por medio de la ecuación de predicción $\hat{Y}_j = b_0 + b_1 X_1 + \ldots + b_{j-1} X_{j-1} + b_{j+1} X_{j+1} + \ldots + b_p X_p$ y el cálculo de los coeficientes de la misma es de la forma $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. La matriz de datos resultante, con tres valores completados por imputación utilizando regresión en la variable X_1 se puede ver en la Tabla 4.4.

Tabla 4.4

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Normal (5, 1)

Método de Imputación por Regresión

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
4.813	3.396	5.569	3.812	5.806
5.726	5.257	4.744	2.798	5.232
4.412	3.944	4.623	5.986	4.010
7.183	6.415	4.704	4.481	6.340
4.864	4.195	3.525	5.327	5.290
5.114	5.529	4.766	5.234	6.479
6.067	5.219	5.118	5.022	6.138
5.059	4.078	5.315	3.996	4.316
4.904	2.829	6.444	4.053	3.708
5.294	4.941	4.649	4.626	4.927
5.294	3.989	5.623	3.814	4.669
3.664	5.615	5.799	3.944	4.156
5.714	5.508	5.941	6.473	5.498
5.714	6.692	4.008	5.056	6.489
4.308	5.591	5.212	3.783	4.454
5.858	4.356	5.238	4.959	4.153
6.254	5.380	3.992	3.872	4.754
3.406	3.991	4.258	3.651	5.663
3.559	4.981	6.082	4.739	4.146
5.571	4.952	4.869	5.954	3.799
4.600	5.000	5.390	5.129	4.880
5.690	4.682	5.088	5.657	4.935
5.816	6.095	4.365	3.832	5.485
5.712	3.126	4.440	4.539	4.405
5.726	5.428	5.444	4.738	4.850
5.669	5.896	4.050	3.787	4.565
6.191	5.731	5.781	3.681	4.921
3.798	5.578	5.569	5.931	6.535
4.980	5.040	4.178	5.011	4.394
4.843	5.677	4.734	4.355	3.653

Elaborado por: G. Cuenca

En la Tabla 4.5 se realiza una comparación entre el valor real y el valor con imputación por la media y regresión.

Tabla 4.5

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

Elaborado por: G. Cuenca

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).

En los Cuadros 4.3, 4.4 y 4.5, podemos apreciar el número de imputaciones sucesivas por medio del *Método de Regresión* que se realiza a los tres datos faltantes en la variable X_1.

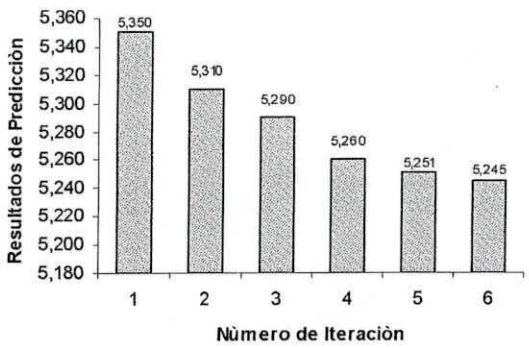
CUADRO 4.3

Efectos de la Imputación en el análisis de datos multivariados
Variables aleatorias independientes con distribución Normal (5,1)
Método de Imputación por Regresión
Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Imputaciones sucesivas para $X_{10,1}$=4.168

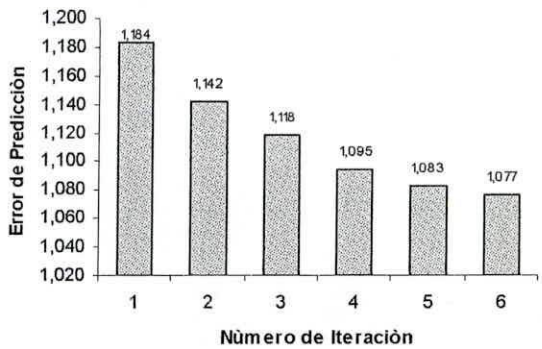
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	5.352	1.184
2	5.310	1.142
3	5.286	1.118
4	5.263	1.095
5	5.251	1.083
6	5.245	1.077

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	6
Media	5.285
Error Estándar	0.017

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	6
Media	1.117
Error Estándar	0.017

Elaborado por: G. Cuenca

En el Cuadro 4.3, se puede ver que el primer resultado de predicción es 5.352 ± 0.017, y el último es 5.245 ± 0.017, donde la media de los resultados de predicción es 5.285 ± 0.017.

CUADRO 4.4

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Normal (5,1)

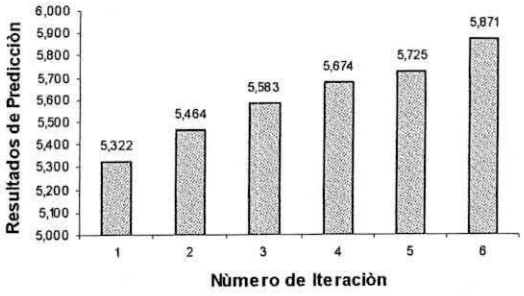
Método de Imputación por Regresión

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Imputaciones sucesivas para $X_{14,1}$=6.629

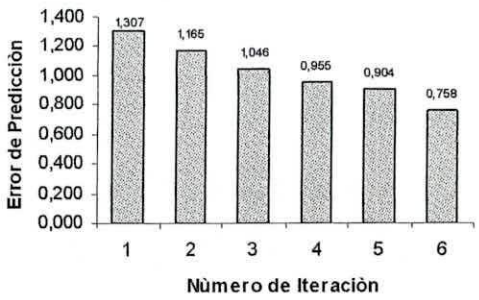
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	5.322	1.307
2	5.464	1.165
3	5.583	1.046
4	5.674	0.955
5	5.725	0.904
6	5.871	0.758

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	6
Media	5.607
Error Estándar	0.080

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	6
Media	1.023
Error Estándar	0.080

Elaborado por: G. Cuenca

En el Cuadro 4.4, se puede ver que el primer resultado de predicción es 5.322 ± 0.080, y el último es 5.871 ± 0.080, donde la media de los resultados de predicción es 5.607 ± 0.080. Mientras que la media del error de predicción es 1.023 ± 0.080.

CUADRO 4.5

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Normal (5,1)

Método de Imputación por Regresión

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

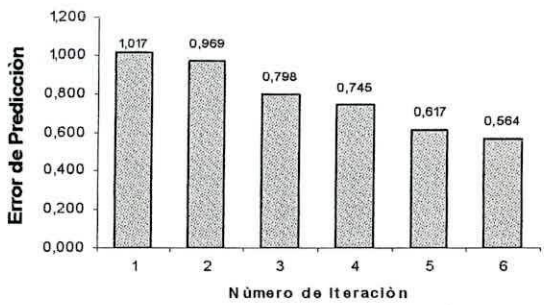
Imputaciones sucesivas para $X_{25,1}$=6.290

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	5.273	1.017
2	5.321	0.969
3	5.492	0.798
4	5.545	0.745
5	5.673	0.617
6	5.726	0.564

Distribución del Resultado de Predicción

Estimadores	Resultado de Predicción
Número de Iteración	6
Media	5.505
Error Estándar	0.075

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	6
Media	0.785
Error Estándar	0.075

Elaborado por: G. Cuenca

En el Cuadro 4.5, se puede observar que el resultado de predicción tiene una media de 5.505 ± 0.075. Se nota también que, en general las imputaciones sucesivas a los tres datos faltantes no tienden al valor observado.

CUADRO 4.6

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Normal (5,1)

Método de Imputación por la Media y Regresión

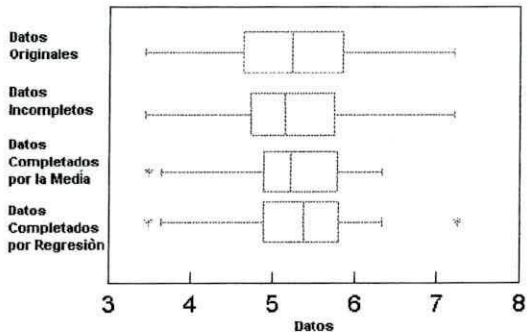
Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_1"

Estimadores

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		30	27	30	30
Media		5,205	5,151	5,151	5,193
Mediana		5,205	5,114	5,151	5,294
Moda		3,406	3,410	5,151	5,294
Varianza		0,891	0,827	0,741	0,763
Desviación Estándar		0,944	0,909	0,861	0,873
Error Estándar		0,172	0,175	0,157	0,159
Coeficiente de Asimetría		-0,161	-0,176	-0,184	-0,314
Curtosis		-0,460	-0,105	0,229	0,102
Rango		3,777	3,780	3,777	3,777
Mínimo		3,406	3,410	3,406	3,406
Máximo		7,183	7,180	7,183	7,183
Percentiles	25	4,553	4,600	4,759	4,759
	50	5,204	5,114	5,151	5,294
	75	5,827	5,726	5,717	5,726

Diagrama de Cajas



Elaborado por: G. Cuenca

Al realizar la imputación por los métodos de "media" y "regresión" se obtuvieron los siguientes resultados (Ver Cuadro 4.6)

El valor de la media de los "datos completados" por *la media* disminuye. comparándolo con los "datos originales" y "datos completados" por *regresión.*

El valor de la varianza de los datos completados por la media disminuye de 0.891 a 0.741, mientras que en los datos completados por regresión este valor se incrementa a 0.763, comparándolo con el valor anterior.

El vector de medias con tres datos completados por la media en X_1 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.151 \\ 4.970 \\ 4.984 \\ 4.608 \\ 4.955 \end{pmatrix}$$

Mientras que el vector de medias con tres datos completados por la regresión en X_1 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.193 \\ 4.970 \\ 4.984 \\ 4.608 \\ 4.955 \end{pmatrix}$$

El efecto que causa en la *matriz de varianzas y covarianzas* y *matriz de correlaciones*, el completar 2% de datos faltantes en una matriz de tamaño 30, por medio de la *imputación por media y regresión*, se presenta en el Cuadro 4.7.

CUADRO 4.7

Efectos de la Imputación en el análisis de datos multivariados
Variables aleatorias independientes con distribución Normal (5,1)
Método de Imputación por la Media y Regresión
Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	0.891				
X_2	0.299	0.891			
X_3	-0.152	-0.138	0.502		
X_4	-0.010	0.034	0.014	0.756	
X_5	0.197	0.315	-0.123	0.090	0.740

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.335	1.000			
X_3	-0.227	-0.206	1.000		
X_4	-0.012	0.042	0.023	1.000	
X_5	0.242	0.388	-0.202	0.120	1.000

Matriz de Varianzas y Covarianzas 10% Datos Completados por Media en "*Variable* X_1"

	X_1	X_2	X_3	X_4	X_5
X_1	**0.741**				
X_2	**0.192**	0.891			
X_3	**-0.132**	-0.138	0.502		
X_4	**-0.037**	0.034	0.014	0.756	
X_5	**0.122**	0.315	-0.123	0.090	0.740

Matriz de Correlaciones 10% Datos Completados por Media en "*Variable* X_1"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	**0.236**	1.000			
X_3	**-0.215**	-0.206	1.000		
X_4	**-0.049**	0.042	0.023	1.000	
X_5	**0.164**	0.388	-0.202	0.120	1.000

Matriz de Varianzas y Covarianzas 10% Datos Completados por Regresión en "*Variable* X_1"

	X_1	X_2	X_3	X_4	X_5
X_1	**0.763**				
X_2	**0.235**	0.891			
X_3	**-0.143**	-0.138	0.502		
X_4	**-0.026**	0.034	0.014	0.756	
X_5	**0.149**	0.315	-0.123	0.090	0.740

Matriz de Correlaciones 10% Datos Completados por Regresiòn en "*Variable* X_1"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	**0.285**	1.000			
X_3	**-0.231**	-0.206	1.000		
X_4	**-0.034**	0.042	0.023	1.000	
X_5	**0.199**	0.388	-0.202	0.120	1.000

Elaborado por: G. Cuenca

Se puede apreciar que los únicos valores que cambian son las covarianzas de la variable X_1 con las demás variables, donde la covarianza entre X_1 y X_2 disminuye de 0.299 a 0.192 en la matriz con 10% de datos completados por la media en la variable X_1, mientras que la

covarianza entre X_1 y X_4 se incrementa en valor absoluto de 0.010 a 0.037.

En la matriz de varianzas y covarianzas de los datos completados por regresión, el valor de las covarianzas de variable X_1 con las demás variables se incrementa, comparándolo con la matriz de varianzas y covarianzas de los datos completados por la media.

Por otro lado, analizando el efecto que causa en la matriz de correlaciones, se nota que la mayor correlación se da entre las variables X_2 y X_5, es decir 0.388, seguida por 0.335 entre las variables X_1 y X_2. En la matriz de correlaciones con 10% de datos completados por la media, la correlación entre X_1 y X_2 disminuye a 0.236, mientras que en la matriz de datos completados por regresión ésta tiene un ligero incremento a 0.285. Se puede apreciar también que en general las variables no están fuertemente correlacionadas entre sí.

4.2.2 Distribución Normal: *Tres datos faltantes, dos* en la variable X_1 y *uno* en la variable X_4 (2% de la matriz), tamaño de muestra n=30

Continuando con la matriz de datos anterior, pero ahora se tienen dos datos faltantes en la variable X_1 y *uno* en la variable X_4, datos cuyas columnas son muestras tomadas de cinco poblaciones todas ellas Normal, independientes e idénticamente distribuidas, con parámetros

μ=5 y σ^2=1. Los datos faltantes son los siguientes: $X_{10,1}$=4.168. $X_{18,4}$=3.651 y el $X_{24,1}$=5.712. Nótese que el 2% de datos faltantes en la matriz, constituye 7% de datos faltantes en la columna correspondiente a X_1 y 3% de datos faltantes en la columna X_4

Tabla 4.6

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Normal (5, 1)

Tamaño de muestra n=30

X_1	X_2	X_3	X_4	X_5
4.813	3.396	5.569	3.812	5.806
5.726	5.257	4.744	2.798	5.232
4.412	3.944	4.623	5.986	4.010
7.183	6.415	4.704	4.481	6.340
4.864	4.195	3.525	5.327	5.290
5.114	5.529	4.766	5.234	6.479
6.067	5.219	**5.118**	5.022	6.138
5.059	4.078	5.315	3.996	4.316
4.904	2.829	6.444	4.053	3.708
4.168	4.941	4.649	4.626	4.927
5.294	3.989	5.623	3.814	4.669
3.664	5.615	5.799	3.944	4.156
5.714	5.508	5.941	6.473	5.498
6.624	6.692	4.008	5.056	6.489
4.308	5.591	5.212	3.783	4.454
5.858	4.356	5.238	4.959	4.153
6.254	5.380	3.992	3.872	4.754
3.406	3.991	4.258	**3.651**	5.663
3.559	4.981	6.082	4.739	4.146
5.571	4.952	4.869	5.954	3.799
4.600	5.000	5.390	5.129	4.880
5.690	4.682	5.088	5.657	4.935
5.816	6.095	4.365	3.832	5.485
5.712	3.126	4.440	4.539	4.405
6.290	5.428	5.444	4.738	4.850
5.669	5.896	4.050	3.787	4.565
6.191	5.731	5.781	3.681	4.921
3.798	5.578	5.569	5.931	6.535
4.980	5.040	4.178	5.011	4.394
4.843	5.677	4.734	4.355	3.653

Elaborado por: G. Cuenca

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 Eliminación por Filas

Como los datos faltantes recayeron en las variables X_1 y X_4, se procede a prescindir de las filas diez, dieciocho y veinticuatro. La matriz de datos resultante con filas eliminadas se muestra en la Tabla 4.7.

Tabla 4.7

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Normal (5,1)

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Matriz de datos con tres filas eliminadas

X_1	X_2	X_3	X_4	X_5
4.813	3.396	5.569	3.812	5.806
5.726	5.257	4.744	2.798	5.232
4.412	3.944	4.623	5.986	4.010
7.183	6.415	4.704	4.481	6.340
4.864	4.195	3.525	5.327	5.290
5.114	5.529	4.766	5.234	6.479
6.067	5.219	5.118	5.022	6.138
5.059	4.078	5.315	3.996	4.316
4.904	2.829	6.444	4.053	3.708
5.294	3.989	5.623	3.814	4.669
3.664	5.615	5.799	3.944	4.156
5.714	5.508	5.941	6.473	5.498
6.624	6.692	4.008	5.056	6.489
4.308	5.591	5.212	3.783	4.454
5.858	4.356	5.238	4.959	4.153
6.254	5.380	3.992	3.872	4.754
3.559	4.981	6.082	4.739	4.146
5.571	4.952	4.869	5.954	3.799
4.600	5.000	5.390	5.129	4.880
5.690	4.682	5.088	5.657	4.935
5.816	6.095	4.365	3.832	5.485
6.290	5.428	5.444	4.738	4.850
5.669	5.896	4.050	3.787	4.565
6.191	5.731	5.781	3.681	4.921
3.798	5.578	5.569	5.931	6.535
4.980	5.040	4.178	5.011	4.394
4.843	5.677	4.734	4.355	3.653

Elaborado por: G. Cuenca

El vector de medias para las veintisiete filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.291 \\ 5.076 \\ 5.043 \\ 4.645 \\ 4.950 \end{pmatrix}$$

CUADRO 4.8

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Normal (5,1)

Método de Eliminación por Filas

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	0.891				
X_2	0.299	0.891			
X_3	-0.152	-0.138	0.502		
X_4	-0.010	0.034	0.014	0.756	
X_5	0.197	0.315	-0.123	0.090	0.740

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.335	1.000			
X_3	-0.227	-0.206	1.000		
X_4	-0.012	0.042	0.023	1.000	
X_5	0.242	0.388	-0.202	0.120	1.000

Matriz de Varianzas y Covarianzas (3 Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	0.811				
X_2	0.291	0.815			
X_3	-0.227	-0.226	0.521		
X_4	-0.078	-0.007	-0.014	0.807	
X_5	0.278	0.339	-0.129	0.125	0.795

Matriz de Correlaciones (3 Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.358	1.000			
X_3	-0.350	-0.348	1.000		
X_4	-0.097	-0.009	-0.022	1.000	
X_5	0.347	0.422	-0.201	0.156	1.000

Elaborado por: G. Cuenca

CUADRO 4.9

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Normal (5,1)

Método de Eliminación por Filas

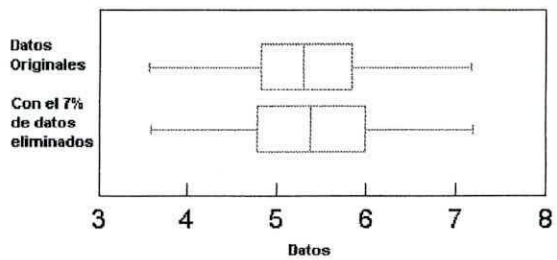
Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_1" y "*Variable* X_4"

Estimadores "*Variable* X_1"

Estimadores		Datos Originales	Con el 7% de datos eliminadas en X_1
n		30	27
Media		5,205	5,291
Mediana		5,205	5,294
Moda		3,410	3,559
Varianza		0,891	0,811
Desviación Estándar		0,944	0,900
Error Estándar		0,172	0,173
Coeficiente de Asimetría		-0,161	-0,133
Curtosis		-0,460	-0,279
Rango		3,780	3,624
Mínimo		3,410	3,559
Máximo		7,180	7,183
Percentiles	25	4,553	4,813
	50	5,204	5,294
	75	5,827	5,858

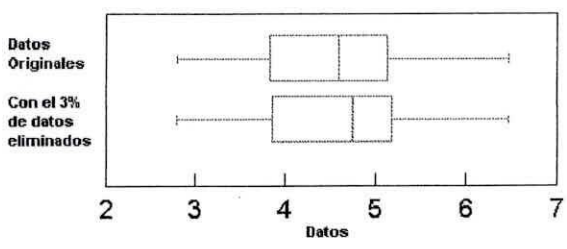
Diagrama de Cajas "*Variable* X_1"



Estimadores "*Variable* X_4"

Estimadores		Datos Originales	Con el 3% de datos eliminadas en X_4
n		30	27
Media		4,608	4,645
Mediana		4,583	4,738
Moda		2,800	2,800
Varianza		0,756	0,807
Desviación Estándar		0,870	0,898
Error Estándar		0,159	0,173
Coeficiente de Asimetría		0,266	0,187
Curtosis		-0,447	-0,589
Rango		3,680	3,680
Mínimo		2,800	2,800
Máximo		6,470	6,470
Percentiles	25	3,828	3,832
	50	4,583	4,738
	75	5,155	5,234

Diagrama de Cajas "*Variable* X_4"



Elaborado por: G. Cuenca

Método de Imputación por la Media y Regresión

La matriz de datos resultante con dos valores completados por imputación por la media en la variable X_1 y uno en X_4, se muestra en la Tabla 4.8.

Tabla 4.8

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Normal (5, 1)

Método de Imputación por la Media

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
4.813	3.396	5.569	3.812	5.806
5.726	5.257	4.744	2.798	5.232
4.412	3.944	4.623	5.986	4.010
7.183	6.415	4.704	4.481	6.340
4.864	4.195	3.525	5.327	5.290
5.114	5.529	4.766	5.234	6.479
6.067	5.219	**5.118**	5.022	6.138
5.059	4.078	5.315	3.996	4.316
4.904	2.829	6.444	4.053	3.708
5.196	4.941	4.649	4.626	4.927
5.294	3.989	5.623	3.814	4.669
3.664	5.615	5.799	3.944	4.156
5.714	5.508	5.941	6.473	5.498
6.624	6.692	4.008	5.056	6.489
4.308	5.591	5.212	3.783	4.454
5.858	4.356	5.238	4.959	4.153
6.254	5.380	3.992	3.872	4.754
3.406	3.991	4.258	**4.641**	5.663
3.559	4.981	6.082	4.739	4.146
5.571	4.952	4.869	5.954	3.799
4.600	5.000	5.390	5.129	4.880
5.690	4.682	5.088	5.657	4.935
5.816	6.095	4.365	3.832	5.485
5.196	3.126	4.440	4.539	4.405
6.290	5.428	5.444	4.738	4.850
5.669	5.896	4.050	3.787	4.565
6.191	5.731	5.781	3.681	4.921
3.798	5.578	5.569	5.931	6.535
4.980	5.040	4.178	5.011	4.394
4.843	5.677	4.734	4.355	3.653

Elaborado por: G. Cuenca

Por otro lado, matriz de datos resultante con dos valores completados por imputación por regresión en la variable X_1 y uno en X_4, se puede ver en la Tabla 4.9.

Tabla 4.9

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Normal (5, 1)

Método de Imputación por Regresión

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
4.813	3.396	5.569	3.812	5.806
5.726	5.257	4.744	2.798	5.232
4.412	3.944	4.623	5.986	4.010
7.183	6.415	4.704	4.481	6.340
4.864	4.195	3.525	5.327	5.290
5.114	5.529	4.766	5.234	6.479
6.067	5.219	**5.118**	5.022	6.138
5.059	4.078	5.315	3.996	4.316
4.904	2.829	6.444	4.053	3.708
3.543	4.941	4.649	4.626	4.927
5.294	3.989	5.623	3.814	4.669
3.664	5.615	5.799	3.944	4.156
5.714	5.508	5.941	6.473	5.498
6.624	6.692	4.008	5.056	6.489
4.308	5.591	5.212	3.783	4.454
5.858	4.356	5.238	4.959	4.153
6.254	5.380	3.992	3.872	4.754
3.406	3.991	4.258	**3.872**	5.663
3.559	4.981	6.082	4.739	4.146
5.571	4.952	4.869	5.954	3.799
4.600	5.000	5.390	5.129	4.880
5.690	4.682	5.088	5.657	4.935
5.816	6.095	4.365	3.832	5.485
5.238	3.126	4.440	4.539	4.405
6.290	5.428	5.444	4.738	4.850
5.669	5.896	4.050	3.787	4.565
6.191	5.731	5.781	3.681	4.921
3.798	5.578	5.569	5.931	6.535
4.980	5.040	4.178	5.011	4.394
4.843	5.677	4.734	4.355	3.653

Elaborado por: G. Cuenca

En la Tabla 4.10 se realiza una comparación entre el valor real y el valor con imputación por la media y regresión. La diferencia en valor absoluto

entre el dato observado de cada variable es menor en el Método de Imputación por Regresión, es decir los datos estimados por medio de la imputación por regresión, están más cercanos a los verdaderos valores, que los de la imputación por la media.

Tabla 4.10

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

Imputación por Media y Regresión en dos valores de la variable X_1

Dato Observado	Resultado de Imputación por la Media	Error |Dato Observado – Resultado de Imputación por Media|	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción|
4.168	5.196	1.028	3.543	0.625
5.712	5.196	0.516	5.238	0.474

Imputación por Media y Regresión en un valor de la variable X_4

Dato Observado	Resultado de Imputación por la Media	Error |Dato Observado – Resultado de Imputación por Media|	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción|
3.651	4.641	0.990	3.872	0.221

Elaborado por: G. Cuenca

En el Cuadro 4.10, se pueden observar los resultados de realizar la imputación por medio de la media y regresión en la variable X_1:

CUADRO 4.10

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Normal (5,1)

Método de Imputación por la Media y Regresión

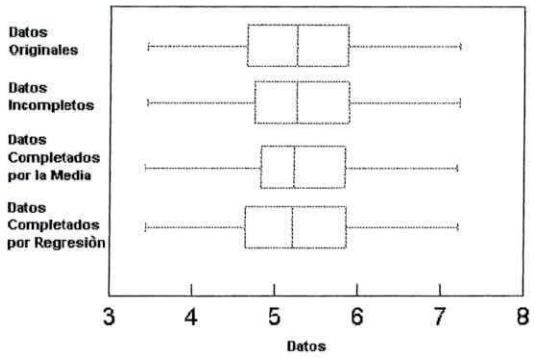
Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_1" y "*Variable* X_4"

Estimadores "*Variable* X_1"

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		30	28	30	30
Media		5,205	5,224	5,224	5,168
Mediana		5,205	5,204	5,196	5,176
Moda		3,410	3,410	5,200	3,410
Varianza		0,891	0,908	0,845	0,939
Desviación Estándar		0,944	0,953	0,919	0,969
Error Estándar		0,172	0,180	0,168	0,177
Coeficiente de Asimetría		-0,161	-0,188	-0,188	-0,174
Curtosis		-0,460	-0,392	-0,196	-0,479
Rango		3,780	3,780	3,780	3,780
Mínimo		3,410	3,410	3,410	3,410
Máximo		7,180	7,180	7,180	7,180
Percentiles	25	4,553	4,653	4,760	4,553
	50	5,204	5,204	5,196	5,176
	75	5,827	5,848	5,827	5,827

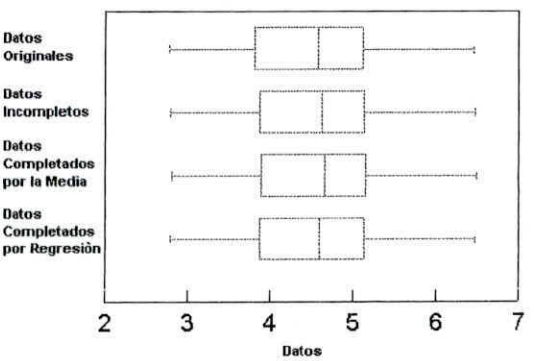
Diagrama de Cajas "*Variable* X_1"



Estimadores "*Variable* X_4"

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		30	29	30	30
Media		4,608	4,641	4,641	4,615
Mediana		4,583	4,626	4,634	4,583
Moda		2,800	2,800	2,800	3,870
Varianza		0,756	0,750	0,724	0,743
Desviación Estándar		0,870	0,866	0,851	0,862
Error Estándar		0,159	0,161	0,155	0,157
Coeficiente de Asimetría		0,266	0,209	0,213	0,274
Curtosis		-0,447	-0,393	-0,297	-0,398
Rango		3,680	3,680	3,680	3,680
Mínimo		2,800	2,800	2,800	2,800
Máximo		6,470	6,470	6,470	6,470
Percentiles	25	3,828	3,852	3,862	3,862
	50	4,583	4,626	4,634	4,583
	75	5,155	5,182	5,155	5,155

Diagrama de Cajas "*Variable* X_4"



Elaborado por: G. Cuenca

El vector de medias con dos datos completados por la media en X_1 y uno en X_4 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.224 \\ 4.970 \\ 4.984 \\ 4.641 \\ 4.955 \end{pmatrix}$$

El vector de medias con dos datos completados por regresión en X_1 y uno en X_4 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.168 \\ 4.970 \\ 4.984 \\ 4.615 \\ 4.955 \end{pmatrix}$$

El efecto que causa en la *matriz de varianzas y covarianzas* y *matriz de correlaciones*, el completar 2% de datos faltantes en una matriz de tamaño 30, por medio de la imputación por media y regresión, se presenta en el Cuadro 4.11.

CUADRO 4.11

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Normal (5,1)

Método de Imputación por la Media y Regresión

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	0.891				
X_2	0.299	0.891			
X_3	-0.152	-0.138	0.502		
X_4	-0.010	0.034	0.014	0.756	
X_5	0.197	0.315	-0.123	0.090	0.740

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.335	1.000			
X_3	-0.227	-0.206	1.000		
X_4	-0.012	0.042	0.023	1.000	
X_5	0.242	**0.388**	-0.202	0.120	1.000

Matriz de Varianzas y Covarianzas 10% Datos Completados por Media en "*Variable* X_1" y "*Variable* X_4"

	X_1	X_2	X_3	X_4	X_5
X_1	**0.845**				
X_2	**0.330**	0.891			
X_3	**-0.154**	-0.138	0.502		
X_4	**-0.070**	**0.001**	**-0.011**	**0.724**	
X_5	**0.205**	0.315	-0.123	**0.114**	0.740

Matriz de Correlaciones 10% Datos Completados por Media en "*Variable* X_1" y "*Variable* X_4"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.381	1.000			
X_3	-0.236	-0.206	1.000		
X_4	-0.089	0.001	-0.018	1.000	
X_5	0.260	0.388	-0.202	0.156	1.000

Matriz de Varianzas y Covarianzas 10% Datos Completados por Regresión en "*Variable* X_1" y "*Variable* X_4"

	X_1	X_2	X_3	X_4	X_5
X_1	**0.939**				
X_2	**0.329**	0.891			
X_3	**-0.136**	-0.138	0.502		
X_4	**-0.022**	**0.027**	**0.009**	**0.743**	
X_5	**0.206**	0.315	-0.123	**0.095**	0.740

Matriz de Correlaciones 10% Datos Completados por Regresión en "*Variable* X_1" y "*Variable* X_4"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.360	1.000			
X_3	-0.197	-0.206	1.000		
X_4	-0.027	0.033	0.014	1.000	
X_5	0.247	0.388	-0.202	0.128	1.000

Elaborado por: G. Cuenca

4.2.3 Distribución Poisson: *Ocho datos faltantes* en una sola variable (5% de la matriz), tamaño de muestra n=30

Se tiene una matriz de datos cuyas columnas son muestras tomadas de cinco poblaciones todas ellas Poisson, independientes e idénticamente distribuìdas, con parámetro $\lambda = 6$, $\mathbf{X} \in \mathrm{M}_{30\times 5}$, *i= 1,2,....30* y *j= 1,2,3,4,5* y se supone que tiene el 5% de datos faltantes, es decir ocho datos, los que

recayeron en la variable X_5 y son: el $X_{3,5}$=6, $X_{7,5}$=3, $X_{10,5}$=3, $X_{14,5}$=4, $X_{18,5}$=5, $X_{21,5}$=5, $X_{25,5}$=9 y el $X_{28,5}$=7.

Nótese que el 5% de datos faltantes en la matriz, constituye 27% de datos faltantes en la columna que corresponde a X_5. (Ver Tabla 4.11)

Los resultados obtenidos para este caso se presentan desde la Tabla 4.11 hasta el Cuadro 4.16

Tabla 4.11

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Poisson $\lambda = 6$

Tamaño de muestra n=30

X_1	X_2	X_3	X_4	X_5
3	10	8	4	2
3	6	7	8	5
6	8	3	10	**6**
6	4	7	10	8
11	5	7	2	4
4	6	9	5	3
9	5	7	6	**3**
3	8	9	6	5
5	2	10	6	8
9	7	4	7	**3**
8	4	7	10	4
3	9	2	8	2
6	9	6	4	4
5	10	6	3	**4**
7	5	6	11	7
5	8	3	5	3
8	11	6	7	8
9	12	7	2	**5**
6	4	8	6	12
5	12	7	9	8
3	2	8	9	**5**
8	9	4	3	10
8	10	4	6	7
4	4	6	7	8
3	8	5	0	**9**
5	9	4	7	11
5	7	8	5	4
4	4	5	2	**7**
8	0	7	6	5
5	8	7	4	9

Elaborado por: G. Cuenca

El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.800 \\ 6.867 \\ 6.233 \\ 5.933 \\ 5.967 \end{pmatrix}$$

Método de Eliminación por Filas

Puesto que los datos faltantes recayeron en la variable X_5 y son: el $X_{3,5}$=6, $X_{7,5}$=3, $X_{10,5}$=3, $X_{14,5}$=4, $X_{18,5}$=5, $X_{25,5}$=9 y el $X_{28,5}$=7, se procede a prescindir de las filas que tienen estos valores "faltantes", donde la matriz de datos resultante con filas eliminadas se muestra en la Tabla 4.12.

Tabla 4.12

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Poisson $\lambda = 6$

Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

Matriz de datos con ocho filas eliminadas

X_1	X_2	X_3	X_4	X_5
3	10	8	4	2
3	6	7	8	5
6	4	7	10	8
11	5	7	2	4
4	6	9	5	3
3	8	9	6	5
5	2	10	6	8
8	4	7	10	4
3	9	2	8	2
6	9	6	4	4
7	5	6	11	7
5	8	3	5	3
8	11	6	7	8
6	4	8	6	12
5	12	7	9	8
8	9	4	3	10
8	10	4	6	7
4	4	6	7	8
5	9	4	7	11
5	7	8	5	4
8	0	7	6	5
5	8	7	4	9

Elaborado por: G. Cuenca

El vector de medias para las veinte y dos filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.727 \\ 6.818 \\ 6.455 \\ 6.318 \\ 6.227 \end{pmatrix}$$

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

Ahora analicemos el efecto que causa en la *matriz de varianzas y covarianzas*, y *matriz de correlaciones*, la eliminación de ocho filas, con un tamaño de muestra *n*=30.(Ver Cuadro 4.12)

CUADRO 4.12

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 6$

Método de Eliminación por Filas

Tamaño de muestra n=30 y 2% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	4.993				
X_2	-0.062	9.361			
X_3	-0.469	-2.140	3.771		
X_4	-0.221	-2.009	-0.156	7.582	
X_5	-0.110	-0.246	-0.061	0.067	7.275

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.009	1.000			
X_3	-0.108	-0.360	1.000		
X_4	-0.036	-0.238	-0.029	1.000	
X_5	-0.018	-0.030	-0.012	0.009	1.000

Matriz de Varianzas y Covarianzas (8 Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	4.494				
X_2	-1.242	9.394			
X_3	-0.537	-2.532	4.069		
X_4	-0.719	-1.082	-0.294	5.465	
X_5	1.208	-0.290	-0.013	0.877	8.374

Matriz de Correlaciones (8 Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.191	1.000			
X_3	-0.126	-0.410	1.000		
X_4	-0.145	-0.151	-0.062	1.000	
X_5	0.197	-0.033	-0.002	0.130	1.000

Elaborado por: G. Cuenca

CUADRO 4.13

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 6$

Método de Eliminación por Filas

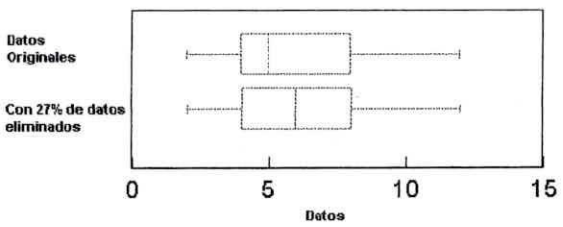
Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_5"

Estimadores

Estimadores		Datos Originales	Con el 27% de datos eliminadas en X_5
n		30	22
Media		5,967	6,227
Mediana		5,000	6,000
Moda		4,000	8,000
Varianza		7,275	8,374
Desviación Estándar		2,697	2,894
Error Estándar		0,492	0,617
Coeficiente de Asimetría		0,456	0,296
Curtosis		-0,633	-0,858
Rango		10,000	10,000
Mínimo		2,000	2,000
Máximo		12,000	12,000
Percentiles	25	4,000	4,000
	50	5,000	6,000

Diagrama de Cajas



Elaborado por: G. Cuenca

Método de Imputación por la Media y Regresión

A continuación se aplica el método de imputación por media y regresión a la misma matriz de datos utilizada en el método de eliminación por filas, es decir se completan datos en la variable X_5 que presenta ocho valores faltantes que son: $X_{3,5}$=6, $X_{7,5}$=3, $X_{10,5}$=3, $X_{14,5}$=4, $X_{18,5}$=5, $X_{25,5}$=9 y el $X_{28,5}$=7. La matriz de datos resultante con ocho valores completados por imputación por la media en la variable X_5 se muestra en la Tabla 4.13.

Tabla 4.13

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Poisson $\lambda = 6$

Método de Imputación por la Media

Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
3	10	8	4	2
3	6	7	8	5
6	8	3	10	**6.227**
6	4	7	10	8
11	5	7	2	4
4	6	9	5	3
9	5	7	6	**6.227**
3	8	9	6	5
5	2	10	6	8
9	7	4	7	**6.227**
8	4	7	10	4
3	9	2	8	2
6	9	6	4	4
5	10	6	3	**6.227**
7	5	6	11	7
5	8	3	5	3
8	11	6	7	8
9	12	7	2	**6.227**
6	4	8	6	12
5	12	7	9	8
3	2	8	9	**6.227**
8	9	4	3	10
8	10	4	6	7
4	4	6	7	8
3	8	5	0	**6.227**
5	9	4	7	11
5	7	8	5	4
4	4	5	2	**6.227**
8	0	7	6	5
5	**8**	**7**	**4**	9

Elaborado por: G. Cuenca

Mientras que la matriz de datos resultante, con ocho valores completados por imputación utilizando regresión en la variable X_5 se puede ver en la Tabla 4.14.

Tabla 4.14

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes que provienen de una distribución Poisson $\lambda = 6$

Método de Imputación por Regresión

Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
3	10	8	4	2
3	6	7	8	5
6	8	3	10	**6.287**
6	4	7	10	8
11	5	7	2	4
4	6	9	5	3
9	5	7	6	**5.110**
3	8	9	6	5
5	2	10	6	8
9	7	4	7	**3.420**
8	4	7	10	4
3	9	2	8	2
6	9	6	4	4
5	10	6	3	**4.310**
7	5	6	11	7
5	8	3	5	3
8	11	6	7	8
9	12	7	2	**6.005**
6	4	8	6	12
5	12	7	9	8
3	2	8	9	**5.106**
8	9	4	3	10
8	10	4	6	7
4	4	6	7	8
3	8	5	0	**8.873**
5	9	4	7	11
5	7	8	5	4
4	4	5	2	**5.517**
8	0	7	6	5
5	8	7	4	9

Elaborado por: G. Cuenca

En la Tabla 4.15 se realiza una comparación entre el dato observado y el dato con imputación por la media y regresión.

Tabla 4.15

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 6$

Comparación de los Métodos de Imputación

Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

27% de datos completados en X_1 por la Media

Dato Observado	Resultado de Imputación por Media	Error |Dato Observado – Resultado de Imputación por Media|
6	6.227	0.227
3	6.227	3.227
3	6.227	3.227
4	6.227	2.227
5	6.227	1.227
5	6.227	1.227
9	6.227	2.773
7	6.227	0.773

27% de datos completados en X_1 por Regresión

Dato Observado	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción
6	6.287	0.287
3	5.110	2.110
3	3.420	0.420
4	4.310	0.310
5	6.005	1.005
5	5.106	0.106
9	8.873	0.127
7	5.517	1.483

Elaborado por: G. Cuenca

Por medio del Cuadro 4.14, podemos apreciar el número de imputaciones sucesivas por medio del Método de Regresión que se realiza a los ocho datos faltantes en la variable X_5.

CUADRO 4.14

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 6$

Método de Imputación por Regresión

Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

Imputaciones sucesivas para $X_{3,5}$=6

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	6.878	0.878
2	6.754	0.754
3	6.632	0.632
4	6.576	0.576
5	6.471	0.471
6	6.323	0.323
7	6.287	0.287

Imputaciones sucesivas para $X_{7,5}$=3

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	6.119	3.119
2	6.032	3.032
3	5.971	2.971
4	5.862	2.862
5	5.531	2.531
6	5.204	2.204
7	5.110	2.110

Imputaciones sucesivas para $X_{10,5}$=3

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	5.429	2.429
2	5.210	2.210
3	4.973	1.973
4	4.415	1.415
5	4.206	1.206
6	3.843	0.843
7	3.420	0.420

Imputaciones sucesivas para $X_{14,5}$=4

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	5.184	1.184
2	5.003	1.003
3	4.852	0.852
4	4.725	0.725
5	4.612	0.612
6	4.561	0.561
7	4.310	0.310

Elaborado por: G. Cuenca

Continúa...

Viene...

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 6$

Método de Imputación por Regresión

Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

Imputaciones sucesivas para $X_{18,5}$=5

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	6.751	1.751
2	6.623	1.623
3	6.541	1.541
4	6.432	1.432
5	6.317	1.317
6	6.210	1.210
7	6.005	1.005

Imputaciones sucesivas para $X_{21,5}$=5

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	5.749	0.749
2	5.663	0.663
3	5.549	0.549
4	5.432	0.432
5	5.316	0.316
6	5.257	0.257
7	5.106	0.106

Imputaciones sucesivas para $X_{25,5}$=9

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	8.215	0.785
2	8.351	0.649
3	8.532	0.468
4	8.673	0.327
5	8.725	0.275
6	8.801	0.199
7	8.873	0.127

Imputaciones sucesivas para $X_{28,5}$=7

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	4.364	2.636
2	4.713	2.287
3	4.846	2.154
4	5.112	1.888
5	5.235	1.765
6	5.418	1.582
7	5.517	1.483

Elaborado por: G. Cuenca

CUADRO 4.15

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 6$

Método de Imputación por la Media y Regresión

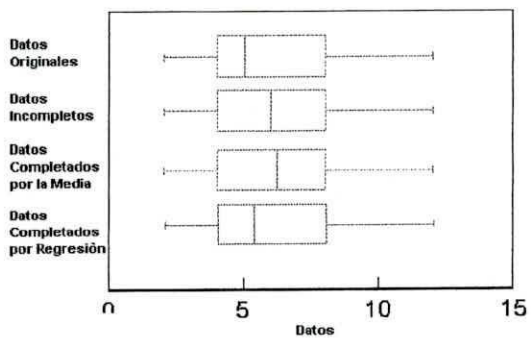
Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_5"

Estimadores

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		30	22	30	30
Media		5,966	6,227	6,227	6,054
Mediana		5,000	6,000	6,227	5,314
Moda		4,000	8,000	6,230	8,000
Varianza		7,275	8,375	6,064	6,779
Desviación Estándar		2,697	2,894	2,462	2,604
Error Estándar		0,492	0,617	0,450	0,475
Coeficiente de Asimetría		0,456	0,296	0,339	0,471
Curtosis		-0,633	-0,858	0,010	-0,429
Rango		10,000	10,000	10,000	10,000
Mínimo		2,000	2,000	2,000	2,000
Máximo		12,000	12,000	12,000	12,000
Percentiles	25	4,000	4,000	4,000	4,000
	50	5,000	6,000	6,227	5,314

Diagrama de Cajas



Elaborado por: G. Cuenca

El vector de medias con ocho "datos completados" por la media en X_5 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.800 \\ 6.867 \\ 6.233 \\ 5.933 \\ 6.227 \end{pmatrix}$$

Mientras que el vector de medias con ocho "datos completados" utilizando regresión en X_5 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 5.800 \\ 6.867 \\ 6.233 \\ 5.933 \\ 6.054 \end{pmatrix}$$

CUADRO 4.16

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 6$

Método de Imputación por la Media y Regresión

Tamaño de muestra n=30 y 5% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	4.993				
X_2	-0.062	9.361			
X_3	-0.469	-2.140	3.771		
X_4	-0.221	-2.009	-0.156	7.582	
X_5	-0.110	-0.246	-0.061	0.067	7.275

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.009	1.000			
X_3	-0.108	-0.360	1.000		
X_4	-0.036	-0.238	-0.029	1.000	
X_5	-0.018	-0.030	-0.012	0.009	1.000

Matriz de Varianzas y Covarianzas 27% Datos Completados por Media en "*Variable* X_5"

	X_1	X_2	X_3	X_4	X_5
X_1	4.993				
X_2	-0.062	9.361			
X_3	-0.469	-2.140	3.771		
X_4	-0.221	-2.009	-0.156	7.582	
X_5	0.875	-0.210	-0.009	0.635	6.064

Matriz de Correlaciones 27% Datos Completados por Media en "*Variable* X_5"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.009	1.000			
X_3	-0.108	-0.360	1.000		
X_4	-0.036	-0.238	-0.029	1.000	
X_5	0.159	-0.028	-0.002	0.094	1.000

Matriz de Varianzas y Covarianzas 27% Datos Completados por Regresión en "*Variable* X_5"

	X_1	X_2	X_3	X_4	X_5
X_1	4.993				
X_2	-0.062	9.361			
X_3	-0.469	-2.140	3.771		
X_4	-0.221	-2.009	-0.156	7.582	
X_5	0.367	-0.033	0.030	0.198	6.779

Matriz de Correlaciones 27% Datos Completados por Regresión en "*Variable* X_5"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.009	1.000			
X_3	-0.108	-0.360	1.000		
X_4	-0.036	-0.238	-0.029	1.000	
X_5	0.063	-0.004	0.006	0.028	1.000

Elaborado por: G. Cuenca

4.2.4 Distribución Exponencial: *Trece datos faltantes* en una sola variable (5% de la matriz), tamaño de muestra n=50

Se tiene una matriz de datos cuyas columnas son muestras tomadas de cinco poblaciones todas ellas Exponencial, independientes e idénticamente distribuìdas, con parámetro $\beta = 2$, $\mathbf{X} \in \mathrm{M}_{50x5}$, *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_2 y son: el $X_{3,2}$=0.335, $X_{6,2}$=2.326, $X_{10,2}$=0.158, $X_{13,2}$=2.019, $X_{18,2}$=1.525, $X_{25,2}$=0.169, $X_{28,2}$=0.606, $X_{31,2}$=4.334, $X_{33,2}$=0.950, $X_{33,2}$=0.950, $X_{37,2}$=4.403, $X_{41,2}$=0.775, $X_{46,2}$=0.337 y $X_{49,2}$=2.209.

Nótese que el 5% de datos faltantes en la matriz, constituye 26% de datos faltantes en la columna que corresponde a X_2.(Ver Tabla 4.16).

Los resultados correspondientes a este caso se presentan desde la Tabla 4.16 hasta el Cuadro 4.19.

Tabla 4.16

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Exponencial β=2

Tamaño de muestra n=50

X_1	X_2	X_3	X_4	X_5
0.308	5.836	3.978	0.967	3.134
0.399	4.329	0.284	1.314	0.790
2.807	**0.335**	2.222	2.019	2.838
0.216	3.516	1.435	0.514	0.656
1.008	6.595	1.681	0.377	1.833
3.936	**2.326**	2.690	2.289	1.863
3.649	0.404	0.034	1.035	1.776
0.249	2.899	0.070	0.492	3.798
0.043	1.064	0.106	6.787	0.260
1.017	**0.158**	1.589	0.309	7.656
2.033	2.217	2.207	0.764	0.158
2.927	3.164	9.718	0.442	0.916
1.598	**2.019**	3.594	0.172	5.409
0.883	0.049	0.377	1.893	0.916
2.811	0.666	0.882	1.502	0.565
1.519	0.882	2.265	4.860	0.524
1.397	0.490	7.271	0.156	0.426
5.182	**1.525**	2.069	0.776	1.220
0.532	2.920	0.592	0.889	1.447
6.186	2.292	2.417	0.008	1.636
3.602	0.475	3.416	1.388	0.935
0.504	0.910	1.758	1.688	5.108
3.137	0.832	0.285	0.375	1.217
2.758	1.241	0.940	4.296	3.130
1.850	**0.169**	5.794	1.026	2.619
1.465	0.496	1.812	3.017	2.464
1.350	2.237	2.395	1.839	0.392
0.941	**0.606**	3.764	2.400	1.776
0.938	1.652	2.348	0.913	1.281
0.018	1.049	11.453	0.113	0.166
2.048	**4.334**	1.654	10.276	1.940
2.575	2.284	0.782	0.405	0.896
0.777	**0.950**	0.390	0.740	3.500
1.352	0.223	0.560	0.038	0.482
2.569	0.074	3.632	5.326	2.012
1.094	7.818	1.188	6.204	0.505
0.390	**4.403**	1.288	0.602	2.145
0.121	1.818	0.168	0.399	0.512
1.622	4.662	1.633	2.688	3.823
1.720	2.455	6.211	0.702	3.818
0.008	**0.775**	0.072	2.432	0.896
0.975	0.041	8.616	4.995	4.742
0.115	1.835	1.188	0.266	0.148
0.184	0.395	0.136	5.116	1.447
0.409	4.056	0.214	0.600	3.625
0.743	**0.337**	0.963	4.158	2.572
2.351	2.916	0.714	1.625	4.066
1.166	5.402	0.126	1.047	7.526
0.903	**2.209**	1.588	0.904	2.928
0.525	1.106	3.467	1.260	0.336

El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 1.538 \\ 2.029 \\ 2.281 \\ 1.889 \\ 2.097 \end{pmatrix}$$

Método de Eliminación por Filas

Puesto que los datos faltantes recayeron en la variable X_2 y son: el $X_{3,2}$=0.335, $X_{6,2}$=2.326, $X_{10,2}$=0.158, $X_{13,2}$=2.019, $X_{18,2}$=1.525, $X_{25,2}$=0.169, $X_{28,2}$=0.606, $X_{31,2}$=4.334, $X_{33,2}$=0.950, $X_{33,2}$=0.950, $X_{37,2}$=4.403, $X_{41,2}$=0.775, $X_{46,2}$=0.337 y $X_{49,2}$=2.209, se procede a prescindir de las filas que tienen estos valores "faltantes", donde la matriz de datos resultante con filas eliminadas se muestra en la Tabla 4.17.

Tabla 4.17

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Exponencial $\beta=2$

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

Matriz de datos con trece filas eliminadas

X_1	X_2	X_3	X_4	X_5
0.308	5.836	3.978	0.967	3.134
0.399	4.329	0.284	1.314	0.790
0.216	3.516	1.435	0.514	0.656
1.008	6.595	1.681	0.377	1.833
3.649	0.404	0.034	1.035	1.776
0.249	2.899	0.070	0.492	3.798
0.043	1.064	0.106	6.787	0.260
2.033	2.217	2.207	0.764	0.158
2.927	3.164	9.718	0.442	0.916
0.883	0.049	0.377	1.893	0.916
2.811	0.666	0.882	1.502	0.565
1.519	0.882	2.265	4.860	0.524
1.397	0.490	7.271	0.156	0.426
0.532	2.920	0.592	0.889	1.447
6.186	2.292	2.417	0.008	1.636
3.602	0.475	3.416	1.388	0.935
0.504	0.910	1.758	1.688	5.108
3.137	0.832	0.285	0.375	1.217
2.758	1.241	0.940	4.296	3.130
1.465	0.496	1.812	3.017	2.464
1.350	2.237	2.395	1.839	0.392
0.938	1.652	2.348	0.913	1.281
0.018	1.049	11.453	0.113	0.166
2.575	2.284	0.782	0.405	0.896
1.352	0.223	0.560	0.038	0.482
2.569	0.074	3.632	5.326	2.012
1.094	7.818	1.188	6.204	0.505
0.121	1.818	0.168	0.399	0.512
1.622	4.662	1.633	2.688	3.823
1.720	2.455	6.211	0.702	3.818
0.975	0.041	8.616	4.995	4.742
0.115	1.835	1.188	0.266	0.148
0.184	0.395	0.136	5.116	1.447
0.409	4.056	0.214	0.600	3.625
2.351	2.916	0.714	1.625	4.066
1.166	5.402	0.126	1.047	7.526
0.525	1.106	3.467	1.260	0.336

Elaborado por: G. Cuenca

El vector de medias para las treinta y siete filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 1.479 \\ 2.197 \\ 2.334 \\ 1.792 \\ 1.823 \end{pmatrix}$$

El efecto que causa en la *matriz de varianzas y covarianzas*, y *matriz de correlaciones*, la eliminación de trece filas, con un tamaño de muestra n=50, se puede ver en el Cuadro 4.17.

CUADRO 4.17

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Exponencial $\beta=2$

Método de Eliminación por Filas

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

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.852				
X_2	-0.355	3.437			
X_3	0.249	-0.651	6.516		
X_4	-0.164	0.072	-0.517	4.472	
X_5	-0.124	0.309	-0.189	-0.225	3.241

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.141	1.000			
X_3	0.072	-0.138	1.000		
X_4	-0.057	0.018	-0.096	1.000	
X_5	-0.050	0.092	-0.041	-0.059	1.000

Matriz de Varianzas y Covarianzas (13 Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	1.769				
X_2	-0.484	3.850			
X_3	0.123	-0.786	8.060		
X_4	-0.279	-0.361	-0.466	3.644	
X_5	-0.034	0.821	-0.264	0.152	3.002

Matriz de Correlaciones (13 Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.186	1.000			
X_3	0.033	-0.141	1.000		
X_4	-0.110	-0.096	-0.086	1.000	
X_5	-0.015	0.242	-0.054	0.046	1.000

Elaborado por: G. Cuenca

Método de Imputación por la Media y Regresión

La matriz de datos resultante con trece valores completados por imputación por la media y regresión en la variable X_2 se muestra en la Tabla 4.18 y 4.19 respectivamente.

Tabla 4.18

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Exponencial $\beta = 2$

Método de Imputación por la Media

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

X_1	X_2	X_3	X_4	X_5
0,308	5,836	3,978	0,967	3,134
0,399	4,329	0,284	1,314	0,790
2,807	**2.197**	2,222	2,019	2,838
0,216	3,516	1,435	0,514	0,656
1,008	6,595	1,681	0,377	1,833
3,936	**2.197**	2,690	2,289	1,863
3,649	0,404	0,034	1,035	1,776
0,249	2,899	0,070	0,492	3,798
0,043	1,064	0,106	6,787	0,260
1,017	**2.197**	1,589	0,309	7,656
2,033	2,217	2,207	0,764	0,158
2,927	3,164	9,718	0,442	0,916
1,598	**2.197**	3,594	0,172	5,409
0,883	0,049	0,377	1,893	0,916
2,811	0,666	0,882	1,502	0,565
1,519	0,882	2,265	4,860	0,524
1,397	0,490	7,271	0,156	0,426
5,182	**2.197**	2,069	0,776	1,220
0,532	2,920	0,592	0,889	1,447
6,186	2,292	2,417	0,008	1,636
3,602	0,475	3,416	1,388	0,935
0,504	0,910	1,758	1,688	5,108
3,137	0,832	0,285	0,375	1,217
2,758	1,241	0,940	4,296	3,130
1,850	**2.197**	5,794	1,026	2,619
1,465	0,496	1,812	3,017	2,464
1,350	2,237	2,395	1,839	0,392
0,941	**2.197**	3,764	2,400	1,776
0,938	1,652	2,348	0,913	1,281
0,018	1,049	11,453	0,113	0,166
2,048	2.197	1,654	10,276	1,940
2,575	2,284	0,782	0,405	0,896
0,777	**2.197**	0,390	0,740	3,500
1,352	0,223	0,560	0,038	0,482
2,569	0,074	3,632	5,326	2,012
1,094	7,818	1,188	6,204	0,505
0,390	**2.197**	1,288	0,602	2,145
0,121	1,818	0,168	0,399	0,512
1,622	4,662	1,633	2,688	3,823
1,720	2,455	6,211	0,702	3,818
0,008	**2.197**	0,072	2,432	0,896
0,975	0,041	8,616	4,995	4,742
0,115	1,835	1,188	0,266	0,148
0,184	0,395	0,136	5,116	1,447
0,409	4,056	0,214	0,600	3,625
0,743	**2.197**	0,963	4,158	2,572
2,351	2,916	0,714	1,625	4,066
1,166	5,402	0,126	1,047	7,526
0,903	**2,197**	1,588	0,904	2,928
0,525	1,106	3,467	1,260	0,336

Tabla 4.19

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias independientes con distribución Exponencial $\beta = 2$

Método de Imputación por Regresión

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

X_1	X_2	X_3	X_4	X_5
0.308	5.836	3.978	0.967	3.134
0.399	4.329	0.284	1.314	0.790
2.807	**2.070**	2.222	2.019	2.838
0.216	3.516	1.435	0.514	0.656
1.008	6.595	1.681	0.377	1.833
3.936	**1.403**	2.690	2.289	1.863
3.649	0.404	0.034	1.035	1.776
0.249	2.899	0.070	0.492	3.798
0.043	1.064	0.106	6.787	0.260
1.017	**3.682**	1.589	0.309	7.656
2.033	2.217	2.207	0.764	0.158
2.927	3.164	9.718	0.442	0.916
1.598	**3.246**	3.594	0.172	5.409
0.883	0.049	0.377	1.893	0.916
2.811	0.666	0.882	1.502	0.565
1.519	0.882	2.265	4.860	0.524
1.397	0.490	7.271	0.156	0.426
5.182	**1.151**	2.069	0.776	1.220
0.532	2.920	0.592	0.889	1.447
6.186	2.292	2.417	0.008	1.636
3.602	0.475	3.416	1.388	0.935
0.504	0.910	1.758	1.688	5.108
3.137	0.832	0.285	0.375	1.217
2.758	1.241	0.940	4.296	3.130
1.850	**1.978**	5.794	1.026	2.619
1.465	0.496	1.812	3.017	2.464
1.350	2.237	2.395	1.839	0.392
0.941	**2.117**	3.764	2.400	1.776
0.938	1.652	2.348	0.913	1.281
0.018	1.049	11.453	0.113	0.166
2.048	**0.907**	1.654	10.276	1.940
2.575	2.284	0.782	0.405	0.896
0.777	**3.181**	0.390	0.740	3.500
1.352	0.223	0.560	0.038	0.482
2.569	0.074	3.632	5.326	2.012
1.094	7.818	1.188	6.204	0.505
0.390	**3.011**	1.288	0.602	2.145
0.121	1.818	0.168	0.399	0.512
1.622	4.662	1.633	2.688	3.823
1.720	2.455	6.211	0.702	3.818
0.008	**2.484**	0.072	2.432	0.896
0.975	0.041	8.616	4.995	4.742
0.115	1.835	1.188	0.266	0.148
0.184	0.395	0.136	5.116	1.447
0.409	4.056	0.214	0.600	3.625
0.743	**2.395**	0.963	4.158	2.572
2.351	2.916	0.714	1.625	4.066
1.166	5.402	0.126	1.047	7.526
0.903	**2.891**	1.588	0.904	2.928
0.525	1.106	3.467	1.260	0.336

En la Tabla 4.20 se realiza una comparación entre el dato observado y el dato con imputación por la media y regresión.

Tabla 4.20

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Exponencial $\beta=2$

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_2 ***por la Media***

Dato Observado	Resultado de Imputación por Media	Error |Dato Observado – Resultado de Imputación por Media|
0.335	2.197	1.862
2.326	2.197	0.129
0.158	2.197	2.039
2.019	2.197	0.178
1.525	2.197	0.672
0.169	2.197	2.028
0.606	2.197	1.591
4.334	2.197	2.137
0.950	2.197	1.247
4.403	2.197	2.206
0.775	2.197	1.422
0.337	2.197	1.860
2.090	2.197	0.107

26% de datos completados en X_2 ***por Regresión***

Dato Observado	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción
0,335	2,070	1,735
2,326	1,403	0,923
0,158	3,682	3,524
2,019	3,246	1,227
1,525	1,151	0,374
0,169	1,978	1,809
0,606	2,117	1,511
4,334	0,907	3,427
0,950	3,181	2,231
4,403	3,011	1,392
0,775	2,484	1,709
0,337	2,395	2,058
2,090	2,891	0,801

Elaborado por: G. Cuenca

CUADRO 4.18

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Exponencial $\beta=2$

Método de Imputación por la Media y Regresión

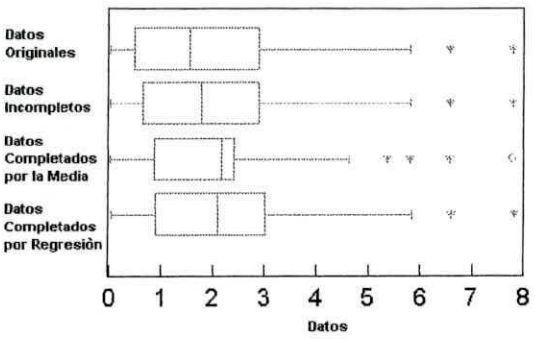
Tamaño de muestra n=50 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_2"

Estimadores

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		50	37	50	50
Media		2,029	2,197	2,197	2,236
Mediana		1,589	1,818	2,197	2,094
Moda		0,040	0,040	2,200	0,040
Varianza		3,437	3,850	2,828	3,011
Desviación Estándar		1,854	1,962	1,682	1,735
Error Estándar		0,262	0,323	0,238	0,245
Coeficiente de Asimetría		1,222	1,164	1,338	1,158
Curtosis		1,115	0,879	2,207	1,487
Rango		7,780	7,780	7,780	7,780
Mínimo		0,040	0,040	0,040	0,040
Máximo		7,820	7,820	7,820	7,820
Percentiles	25	0,495	0,581	0,903	0,901
	50	1,589	1,818	2,197	2,094
	75	2,917	3,042	2,566	3,049

Diagrama de Cajas



Elaborado por: G. Cuenca

El vector de medias con trece datos completados por la media en X_2 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 1.538 \\ 2.197 \\ 2.281 \\ 1.889 \\ 2.097 \end{pmatrix}$$

Mientras que el vector de medias con trece datos completados por la regresión en X_2 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 1.538 \\ 2.236 \\ 2.281 \\ 1.889 \\ 2.097 \end{pmatrix}$$

El efecto que causa en la *matriz de varianzas* y *covarianzas* y *matriz de correlaciones*, el completar 5% de datos faltantes en una matriz de tamaño 50, por medio de la imputación por media y regresión, se presenta en el Cuadro 4.19.

CUADRO 4.19

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Exponencial $\beta=2$

Método de Imputación por la Media y Regresión

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

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.852				
X_2	-0.355	3.437			
X_3	0.249	-0.651	6.516		
X_4	-0.164	0.072	-0.517	4.472	
X_5	-0.124	0.309	-0.189	-0.225	3.241

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.141	1.000			
X_3	0.072	-0.138	1.000		
X_4	-0.057	0.018	-0.096	1.000	
X_5	-0.050	0.092	-0.041	-0.059	1.000

Matriz de Varianzas y Covarianzas 26% Datos Completados por Media en "*Variable* X_2"

	X_1	X_2	X_3	X_4	X_5
X_1	1.852				
X_2	-0.356	2.828			
X_3	0.249	-0.578	6.516		
X_4	-0.164	-0.265	-0.517	4.472	
X_5	-0.124	0.603	-0.189	-0.225	3.241

Matriz de Correlaciones 26% Datos Completados por Media en "*Variable* X_2"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.155	1.000			
X_3	0.072	-0.135	1.000		
X_4	-0.057	-0.075	-0.096	1.000	
X_5	-0.050	0.199	-0.041	-0.059	1.000

Matriz de Varianzas y Covarianzas 26% Datos Completados por Regresión en "*Variable* X_2"

	X_1	X_2	X_3	X_4	X_5
X_1	1.852				
X_2	-0.560	3.011			
X_3	0.249	-0.657	6.516		
X_4	-0.164	-0.597	-0.517	4.472	
X_5	-0.124	0.901	-0.189	-0.225	3.241

Matriz de Correlaciones 26% Datos Completados por Regresión en "*Variable* X_2"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	-0.237	1.000			
X_3	0.072	-0.148	1.000		
X_4	-0.057	-0.163	-0.096	1.000	
X_5	-0.050	0.289	-0.041	-0.059	1.000

Elaborado por: G. Cuenca

4.3 Matrices de Datos con variables aleatorias dependientes

En esta sección se realiza la comparación de los Métodos de Imputación, utilizando matrices de datos con variables aleatorias dependientes, con las distribuciones Normal, Poisson y Exponencial.

4.3.1 Distribución Normal: *Trece datos faltantes* en una sola variable (5% de la matriz), tamaño de muestra n=50

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, $\mathbf{X} \in \mathrm{M}_{50\mathrm{x}5}$, *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.(Ver Tabla 4.21)

Tabla 4.21

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Normal (10, 1)

Tamaño de muestra n=50

X_1	X_2	X_3	X_4	X_5
10.795	10.399	10.777	10.610	11.217
9.866	9.975	**9.010**	9.863	10.929
7.841	7.267	8.513	8.214	8.712
11.869	10.340	11.380	10.312	11.007
10.350	12.547	**11.221**	10.324	10.532
9.299	10.392	**10.102**	10.320	9.449
10.534	9.264	10.164	9.067	9.447
10.325	11.979	11.486	10.526	11.554
10.288	10.920	**9.927**	9.554	11.840
9.232	9.984	10.538	9.633	9.045
11.463	10.285	**10.718**	9.156	9.243
9.427	10.861	9.573	9.717	8.939
10.678	9.843	10.905	10.302	9.628
9.580	9.948	9.478	10.324	9.885
9.714	9.214	9.334	10.042	9.996
8.282	8.433	9.356	9.677	8.955
11.562	10.166	**11.504**	10.953	10.491
9.588	10.713	10.476	11.278	11.123
9.649	10.292	9.565	10.365	9.811
10.100	10.191	9.732	10.977	9.444
12.278	11.190	**12.263**	10.723	11.435
9.723	11.318	11.123	11.680	10.760
10.240	9.289	**10.329**	9.904	9.946
9.526	9.516	11.707	10.888	10.849
9.059	9.980	8.240	10.071	10.326
8.777	9.674	9.730	10.410	9.548
10.328	10.406	10.584	10.678	10.698
10.047	9.038	9.562	9.427	9.446
10.290	9.460	**10.655**	9.544	9.785
9.312	10.242	9.415	10.194	9.982
9.330	8.964	9.607	9.561	9.740
9.819	9.472	**9.547**	9.324	9.188
9.774	9.301	10.327	10.016	9.132
9.706	9.902	10.165	10.196	10.329
9.645	9.857	10.916	10.587	9.147
11.296	11.196	10.420	10.252	10.928
9.854	9.483	**9.509**	9.731	10.447
9.163	9.153	11.430	10.506	10.708
9.435	9.901	9.737	10.184	10.011
10.232	9.714	9.208	9.834	9.961
9.658	8.187	**9.189**	8.847	9.840
9.695	9.276	10.903	10.868	10.161
11.174	12.345	11.321	11.366	11.804
9.630	11.485	11.574	12.158	11.666
9.131	10.067	9.754	9.340	9.765
10.164	9.141	**9.549**	9.524	10.820
9.455	10.444	9.792	10.016	10.999
10.790	9.637	9.035	9.795	9.584
11.428	10.079	11.551	10.164	10.742
10.463	9.852	9.813	9.842	10.429

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 Eliminación por Filas

Debido a que los datos faltantes 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, se procede a prescindir de las filas que tienen estos valores "faltantes", donde la matriz de datos resultante con filas eliminadas se muestra en la Tabla 4.22.

Tabla 4.22

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Normal (10, 1)

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

Matriz de datos con trece filas eliminadas

X_1	X_2	X_3	X_4	X_5
10.795	10.399	10.777	10.610	11.217
7.841	7.267	8.513	8.214	8.712
11.869	10.340	11.380	10.312	11.007
10.534	9.264	10.164	9.067	9.447
10.325	11.979	11.486	10.526	11.554
9.232	9.984	10.538	9.633	9.045
9.427	10.861	9.573	9.717	8.939
10.678	9.843	10.905	10.302	9.628
9.580	9.948	9.478	10.324	9.885
9.714	9.214	9.334	10.042	9.996
8.282	8.433	9.356	9.677	8.955
9.588	10.713	10.476	11.278	11.123
9.649	10.292	9.565	10.365	9.811
10.100	10.191	9.732	10.977	9.444
9.723	11.318	11.123	11.680	10.760
9.526	9.516	11.707	10.888	10.849
9.059	9.980	8.240	10.071	10.326
8.777	9.674	9.730	10.410	9.548
10.328	10.406	10.584	10.678	10.698
10.047	9.038	9.562	9.427	9.446
9.312	10.242	9.415	10.194	9.982
9.330	8.964	9.607	9.561	9.740
9.774	9.301	10.327	10.016	9.132
9.706	9.902	10.165	10.196	10.329
9.645	9.857	10.916	10.587	9.147
11.296	11.196	10.420	10.252	10.928
9.163	9.153	11.430	10.506	10.708
9.435	9.901	9.737	10.184	10.011
10.232	9.714	9.208	9.834	9.961
9.695	9.276	10.903	10.868	10.161
11.174	12.345	11.321	11.366	11.804
9.630	11.485	11.574	12.158	11.666
9.131	10.067	9.754	9.340	9.765
9.455	10.444	9.792	10.016	10.999
10.790	9.637	9.035	9.795	9.584
11.428	10.079	11.551	10.164	10.742
10.463	9.852	9.813	9.842	10.429

Elaborado por: G. Cuenca

El vector de medias para las treinta y siete filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 9.858 \\ 10.002 \\ 10.194 \\ 10.245 \\ 10.148 \end{pmatrix}$$

El vector de medias de los datos originales y de los datos con filas eliminadas no coincide.

Ahora analicemos en el Cuadro 4.20, el efecto que causa en la *matriz de varianzas y covarianzas*, y *matriz de correlaciones*, la eliminación de trece filas, con un tamaño de muestra *n*=50.

CUADRO 4.20

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10, 1)

Método de Eliminación por Filas

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

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	0.758				
X_2	0.387	0.953			
X_3	0.439	0.465	0.828		
X_4	0.135	0.439	0.396	0.517	
X_5	0.317	0.483	0.363	0.327	0.668

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.455	1.000			
X_3	0.554	0.524	1.000		
X_4	0.215	0.625	0.606	1.000	
X_5	0.445	0.606	0.488	0.556	1.000

Matriz de Varianzas y Covarianzas (Trece Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	0.711				
X_2	0.399	0.898			
X_3	0.357	0.414	0.812		
X_4	0.163	0.470	0.411	0.533	
X_5	0.338	0.540	0.445	0.401	0.678

Matriz de Correlaciones (Trece Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.499	1.000			
X_3	0.470	0.484	1.000		
X_4	0.266	0.680	0.625	1.000	
X_5	0.487	0.693	0.600	0.667	1.000

Elaborado por: G. Cuenca

Se puede apreciar que la mayor covarianza en la matriz de datos originales se da entre las variables X_2 y X_5 es decir 0.483; mientras que en la matriz con tres filas eliminadas este valor aumenta a 0.540.

En la matriz de correlaciones de datos originales, la mayor correlación se da entre las variables X_2 y X_4, es decir 0.625, cuyo valor se incrementa a 0.680 en la matriz de correlaciones con trece filas eliminadas. Se puede apreciar también que la mayor correlación en la matriz de datos con trece filas eliminadas se da entre las variables X_2 y X_5, es decir 0.693. En general, se puede decir que las variables tienen una correlación fuerte.

También se realiza el análisis de la variable que presenta datos faltantes, en este caso la variable X_3. (Ver Cuadro 4.21)

CUADRO 4.21

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

Método de Eliminación por Filas

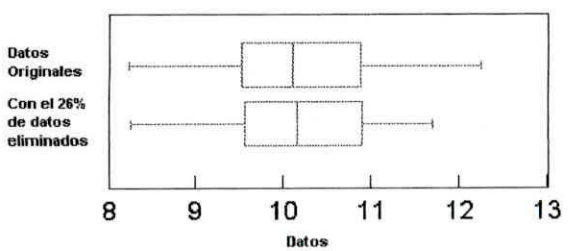
Tamaño de muestra n=50 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_3"

Estimadores

Estimadores		Datos Originales	Con el 26% de datos eliminadas en X_3
n		50	37
Media		10,214	10,194
Mediana		10,133	10,164
Moda		8,240	8,240
Varianza		0,828	0,812
Desviación Estándar		0,910	0,901
Error Estándar		0,129	0,148
Coeficiente de Asimetría		0,173	-0,012
Curtosis		-0,639	-0,759
Rango		4,020	3,470
Mínimo		8,240	8,240
Máximo		12,260	11,710
Percentiles	25	9,549	9,564
	50	10,133	10,164
	75	10,907	10,911

Diagrama de Cajas



Elaborado por: G. Cuenca

En el Cuadro 4.21, podemos apreciar que con el 26% de datos eliminados en la tercera columna de la matriz de datos (Variable X_3), el valor de la media aumentó de 10.214 a 10.194 , la varianza disminuyó de 0.828 a 0.812.

Método de Imputación por la Media y Regresión

A continuación se aplica el método de imputación por media y regresión a la misma matriz de datos utilizada en el método de eliminación por filas, es decir se completan datos en la variable X_3 que presenta trece valores faltantes que 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.

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 $X_{46,3}$.

La matriz de datos resultante con trece valores completados por *imputación por la media* y *regresión* en la variable X_3 se muestra en la Tabla 4.23 y 4.24 respectivamente.

Tabla 4.23

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Normal (10, 1)

Método de Imputación por la Media

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

X_1	X_2	X_3	X_4	X_5
10.795	10.399	10.777	10.610	11.217
9.866	9.975	**10.194**	9.863	10.929
7.841	7.267	8.513	8.214	8.712
11.869	10.340	11.380	10.312	11.007
10.350	12.547	**10.194**	10.324	10.532
9.299	10.392	**10.194**	10.320	9.449
10.534	9.264	10.164	9.067	9.447
10.325	11.979	11.486	10.526	11.554
10.288	10.920	**10.194**	9.554	11.840
9.232	9.984	10.538	9.633	9.045
11.463	10.285	**10.194**	9.156	9.243
9.427	10.861	9.573	9.717	8.939
10.678	9.843	10.905	10.302	9.628
9.580	9.948	9.478	10.324	9.885
9.714	9.214	9.334	10.042	9.996
8.282	8.433	9.356	9.677	8.955
11.562	10.166	**10.194**	10.953	10.491
9.588	10.713	10.476	11.278	11.123
9.649	10.292	9.565	10.365	9.811
10.100	10.191	9.732	10.977	9.444
12.278	11.190	**10.194**	10.723	11.435
9.723	11.318	11.123	11.680	10.760
10.240	9.289	**10.194**	9.904	9.946
9.526	9.516	11.707	10.888	10.849
9.059	9.980	8.240	10.071	10.326
8.777	9.674	9.730	10.410	9.548
10.328	10.406	10.584	10.678	10.698
10.047	9.038	9.562	9.427	9.446
10.290	9.460	**10.194**	9.544	9.785
9.312	10.242	9.415	10.194	9.982
9.330	8.964	9.607	9.561	9.740
9.819	9.472	**10.194**	9.324	9.188
9.774	9.301	10.327	10.016	9.132
9.706	9.902	10.165	10.196	10.329
9.645	9.857	10.916	10.587	9.147
11.296	11.196	10.420	10.252	10.928
9.854	9.483	**10.194**	9.731	10.447
9.163	9.153	11.430	10.506	10.708
9.435	9.901	9.737	10.184	10.011
10.232	9.714	9.208	9.834	9.961
9.658	8.187	**10.194**	8.847	9.840
9.695	9.276	10.903	10.868	10.161
11.174	12.345	11.321	11.366	11.804
9.630	11.485	11.574	12.158	11.666
9.131	10.067	9.754	9.340	9.765
10.164	9.141	**10.194**	9.524	10.820
9.455	10.444	9.792	10.016	10.999
10.790	9.637	9.035	9.795	9.584
11.428	10.079	11.551	10.164	10.742
10.463	9.852	9.813	9.842	10.429

Tabla 4.24

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Normal (10, 1)

Método de Imputación por Regresión

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

X_1	X_2	X_3	X_4	X_5
10.795	10.399	10.777	10.610	11.217
9.866	9.975	**9.110**	9.863	10.929
7.841	7.267	8.513	8.214	8.712
11.869	10.340	11.380	10.312	11.007
10.350	12.547	**11.215**	10.324	10.532
9.299	10.392	**10.112**	10.320	9.449
10.534	9.264	10.164	9.067	9.447
10.325	11.979	11.486	10.526	11.554
10.288	10.920	**9.953**	9.554	11.840
9.232	9.984	10.538	9.633	9.045
11.463	10.285	**10.709**	9.156	9.243
9.427	10.861	9.573	9.717	8.939
10.678	9.843	10.905	10.302	9.628
9.580	9.948	9.478	10.324	9.885
9.714	9.214	9.334	10.042	9.996
8.282	8.433	9.356	9.677	8.955
11.562	10.166	**11.510**	10.953	10.491
9.588	10.713	10.476	11.278	11.123
9.649	10.292	9.565	10.365	9.811
10.100	10.191	9.732	10.977	9.444
12.278	11.190	**12.253**	10.723	11.435
9.723	11.318	11.123	11.680	10.760
10.240	9.289	**10.333**	9.904	9.946
9.526	9.516	11.707	10.888	10.849
9.059	9.980	8.240	10.071	10.326
8.777	9.674	9.730	10.410	9.548
10.328	10.406	10.584	10.678	10.698
10.047	9.038	9.562	9.427	9.446
10.290	9.460	**10.652**	9.544	9.785
9.312	10.242	9.415	10.194	9.982
9.330	8.964	9.607	9.561	9.740
9.819	9.472	**9.545**	9.324	9.188
9.774	9.301	10.327	10.016	9.132
9.706	9.902	10.165	10.196	10.329
9.645	9.857	10.916	10.587	9.147
11.296	11.196	10.420	10.252	10.928
9.854	9.483	**9.507**	9.731	10.447
9.163	9.153	11.430	10.506	10.708
9.435	9.901	9.737	10.184	10.011
10.232	9.714	9.208	9.834	9.961
9.658	8.187	**9.181**	8.847	9.840
9.695	9.276	10.903	10.868	10.161
11.174	12.345	11.321	11.366	11.804
9.630	11.485	11.574	12.158	11.666
9.131	10.067	9.754	9.340	9.765
10.164	9.141	**9.539**	9.524	10.820
9.455	10.444	9.792	10.016	10.999
10.790	9.637	9.035	9.795	9.584
11.428	10.079	11.551	10.164	10.742
10.463	9.852	9.813	9.842	10.429

Tabla 4.25

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

Elaborado por: G. Cuenca

Se puede notar, por medio de la Tabla 4.25 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*.

CUADRO 4.22

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

Método de Imputación por Regresión

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

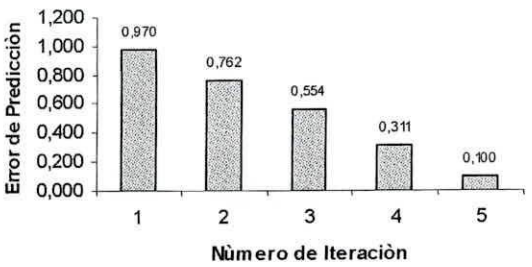
Imputaciones sucesivas para $X_{2,3}$=9.010

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	9.980	0,970
2	9.772	0,762
3	9.564	0,554
4	9.321	0,311
5	9.110	0,100

Distribución del Resultado de Predicción

Estimadores	Resultado de Predicción
Número de Iteración	5
Media	9.549
Error Estándar	0.155

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.539
Error Estándar	0.155

Elaborado por: G. Cuenca

En el Cuadro 4.22, se puede ver que el primer resultado de predicción es 9.980 ± 0.155, y el último es 9.110 ± 0.155, donde la media de los resultados de predicción es 9.549 ± 0.155.

CUADRO 4.23

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

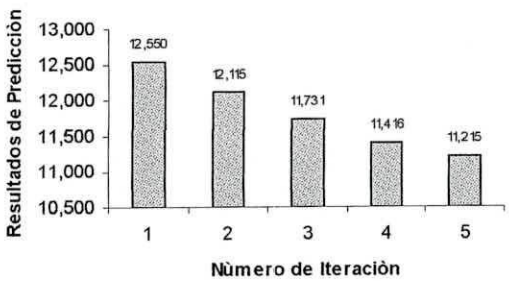
Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{5,3}$=11.221

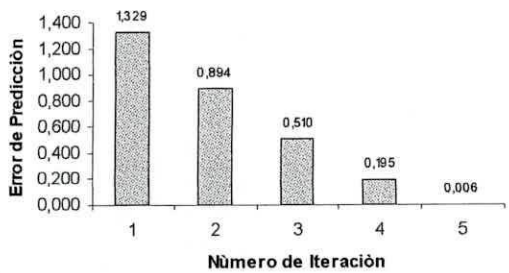
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	12.550	1,329
2	12.115	0,894
3	11.731	0,510
4	11.416	0,195
5	11.215	0,006

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	11.805
Error Estándar	0.240

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.587
Error Estándar	0.239

Elaborado por: G. Cuenca

En el Cuadro 4.23, se puede ver que el primer dato resultado de predicción es 12.550 ± 0.240, y el último es 11.215 ± 0.240, donde la media de los resultados de predicción es 11.805 ± 0.240 y la media del error de predicción es 0.587 ± 0.240.

CUADRO 4.24

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

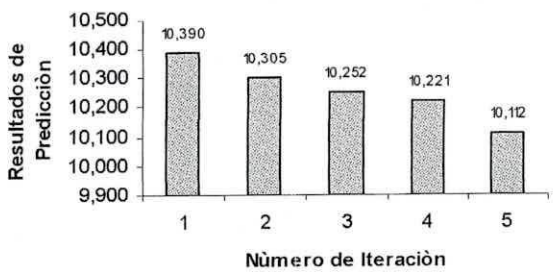
Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{6,3}$=10.102

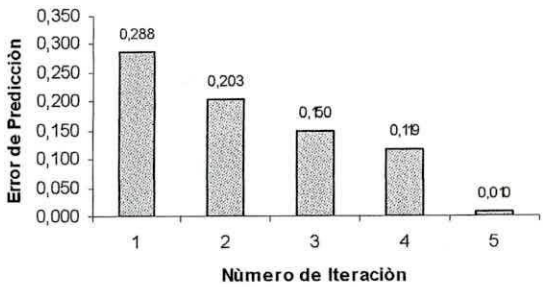
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	10.390	0,288
2	10.305	0,203
3	10.252	0,150
4	10.221	0,119
5	10.112	0,010

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	10.256
Error Estándar	0.046

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.154
Error Estándar	0.046

Elaborado por: G. Cuenca

El Cuadro 4.24, nos muestra que el primer resultado de predicción es 10.390 ± 0.046, y el último es 10.112 ± 0.046, donde la media de los resultados de predicción es 10.256 ± 0.046 y la media del error de predicción es 0.154 ± 0.046.

CUADRO 4.25

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

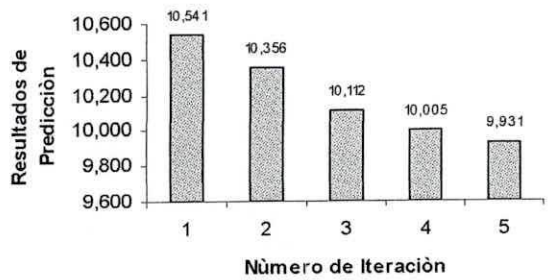
Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{9,3}$=9.927

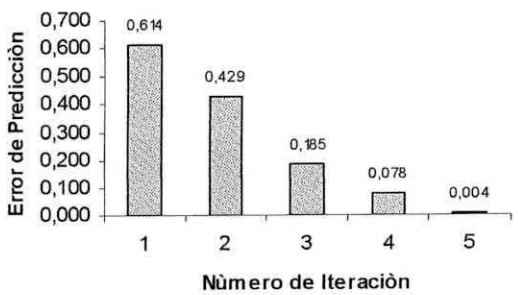
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	10.541	0,614
2	10.356	0,429
3	10.112	0,185
4	10.005	0,078
5	9.931	0,004

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	10.189
Error Estándar	0.114

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.262
Error Estándar	0.114

Elaborado por: G. Cuenca

El Cuadro 4.25, nos muestra que el primer resultado de predicción es 10.541 ± 0.114, y el último es 9.931 ± 0.114, donde la media de los resultados de predicción es 10.189 ± 0.114 y la media del error de predicción es 0.262 ± 0.114.

CUADRO 4.26

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

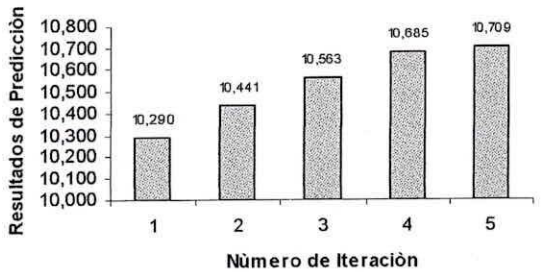
Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{11,3}$=10.718

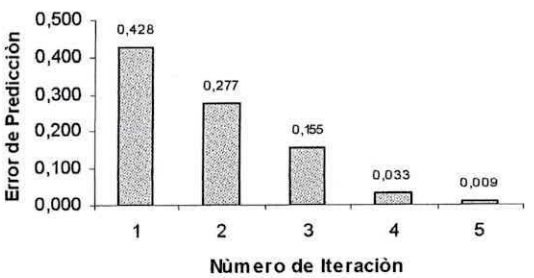
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	10.290	0,428
2	10.441	0,277
3	10.563	0,155
4	10.685	0,033
5	10.709	0,009

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	10.538
Error Estándar	0.078

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.184
Error Estándar	0.078

Elaborado por: G. Cuenca

El Cuadro 4.26, nos muestra que el primer resultado de predicción es 10.290 ± 0.078, y el último es 10.709 ± 0.078, donde la media de los resultados de predicción es 10.538 ± 0.078 y la media del error de predicción es 0.184 ± 0.078.

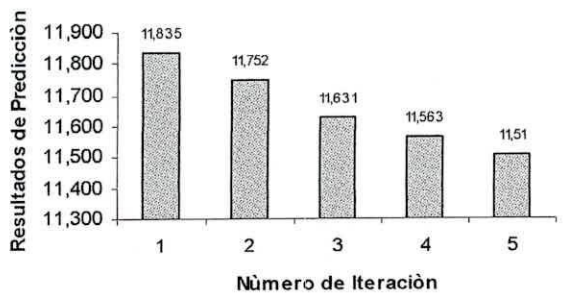
CUADRO 4.27

Efectos de la Imputación en el análisis de datos multivariados
Variables aleatorias dependientes con distribución Normal (10,1)
Método de Imputación por Regresión
Tamaño de muestra n=50 y 5% de datos faltantes en la matriz

Imputaciones sucesivas para $X_{17,3}$=11.504

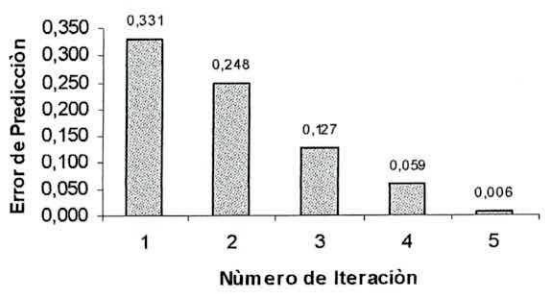
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	11.835	0,331
2	11.752	0,248
3	11.631	0,127
4	11.563	0,059
5	11.510	0,006

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	11.658
Error Estándar	0.060

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.154
Error Estándar	0.060

Elaborado por: G. Cuenca

El Cuadro 4.27, nos muestra que el primer resultado de predicción es 11.835 ± 0.060, y el último es 11.510 ± 0.060, donde la media de los resultados de predicción es 11.658 ± 0.060 y la media del error de predicción es 0.154 ± 0.060.

CUADRO 4.28

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

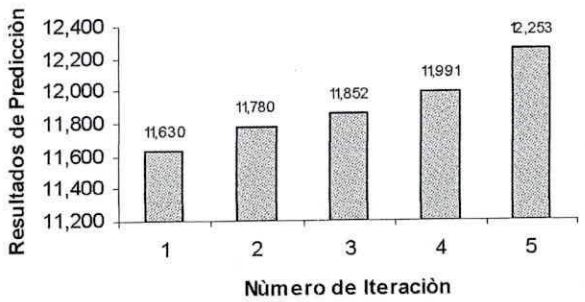
Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{21,3}$=12.263

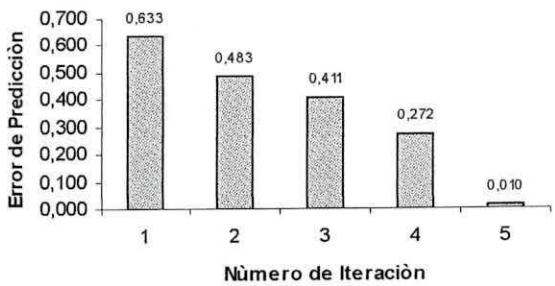
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	11.630	0,633
2	11.780	0,483
3	11.852	0,411
4	11.991	0,272
5	12.253	0,010

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	11.901
Error Estándar	0.105

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.362
Error Estándar	0.105

Elaborado por: G. Cuenca

El Cuadro 4.28, nos muestra que el primer resultado de predicción es 11.630 ± 0.150, y el último es 12.253 ± 0.150, donde la media de los resultados de predicción es 11.901 ± 0.150 y la media del error de predicción es 0.362 ± 0.0105.

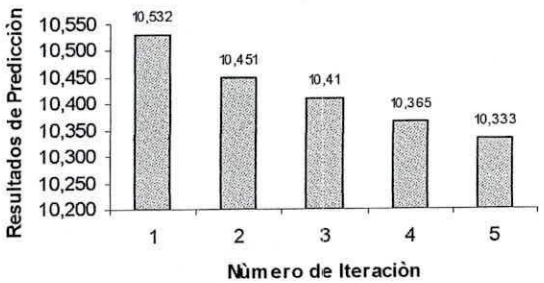
CUADRO 4.29

Efectos de la Imputación en el análisis de datos multivariados
Variables aleatorias dependientes con distribución Normal (10,1)
Método de Imputación por Regresión
Tamaño de muestra n=50 y 5% de datos faltantes en la matriz

Imputaciones sucesivas para $X_{23,3}$=10.329

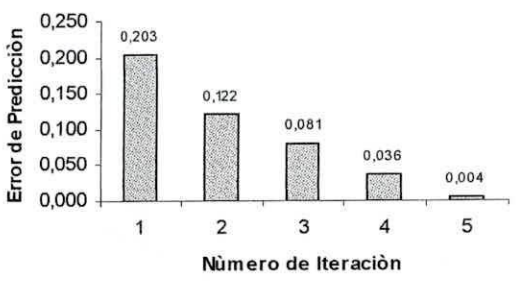
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	10.532	0,203
2	10.451	0,122
3	10.410	0,081
4	10.365	0,036
5	10.333	0,004

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	10.418
Error Estándar	0.035

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.089
Error Estándar	0.035

Elaborado por: G. Cuenca

El Cuadro 4.29, nos muestra que el primer resultado de predicción es 10.532 ± 0.035, y el último es 10.333 ± 0.035, donde la media de los resultados de predicción es 10.418 ± 0.035 y la media del error de predicción es 0.089 ± 0.035.

CUADRO 4.30

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

Método de Imputación por Regresión

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

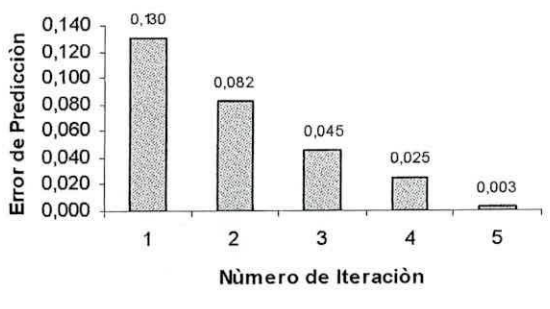
Imputaciones sucesivas para $X_{29,3}$=10.655

Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	10.525	0,130
2	10.573	0,082
3	10.610	0,045
4	10.630	0,025
5	10.652	0,003

Distribución del Resultado de Predicción

Estimadores	Resultado de Predicción
Número de Iteración	5
Media	10.598
Error Estándar	0.022

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.026
Error Estándar	0.022

Elaborado por: G. Cuenca

El Cuadro 4.30, nos muestra que el primer resultado de predicción es 10.525 ± 0.022, y el último es 10.652 ± 0.022, donde la media de los resultados de predicción es 10.598 ± 0.022 y la media del error de predicción es 0.026 ± 0.022.

CUADRO 4.31

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

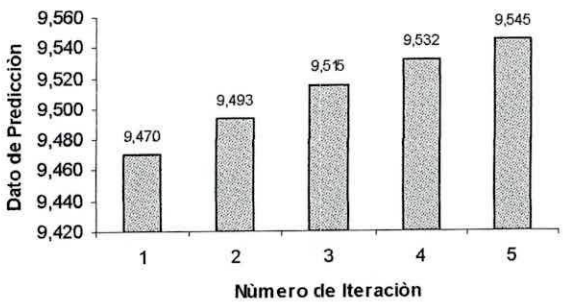
Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{32,3}$=9.547

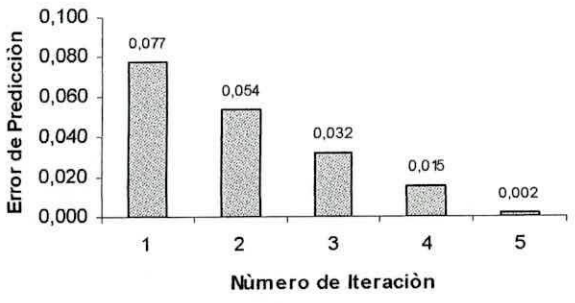
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	9.470	0,077
2	9.493	0,054
3	9.515	0,032
4	9.532	0,015
5	9.545	0,002

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	9.511
Error Estándar	0.013

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.036
Error Estándar	0.013

Elaborado por: G. Cuenca

El Cuadro 4.31, nos muestra que el primer resultado de predicción es 9.470 ± 0.013, y el último es 9.545 ± 0.013, donde la media de los resultados de predicción es 9.511 ± 0.013 y la media del error de predicción es 0.036 ± 0.013.

CUADRO 4.32

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

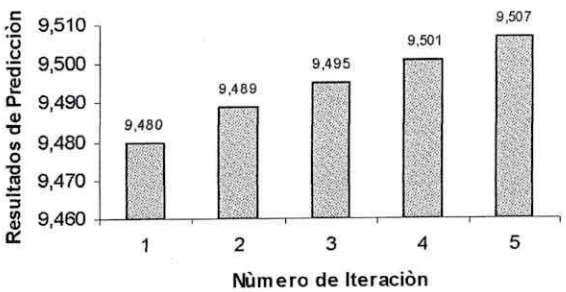
Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{37,3}$=9.509

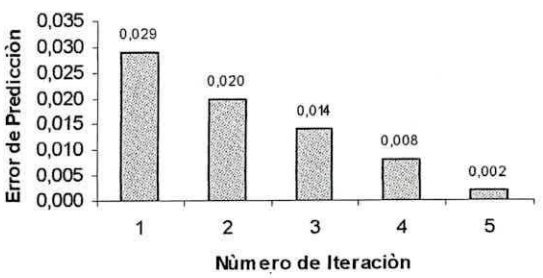
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	9.480	0,029
2	9.489	0,020
3	9.495	0,014
4	9.501	0,008
5	9.507	0,002

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	9.494
Error Estándar	0.005

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.015
Error Estándar	0.005

Elaborado por: G. Cuenca

El Cuadro 4.32, nos muestra que el primer resultado de predicción es 9.480 ± 0.005, y el último es 9.507 ± 0.005, donde la media de los resultados de predicción es 9.494 ± 0.005 y la media del error de predicción es 0.015±0.005.

CUADRO 4.33

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

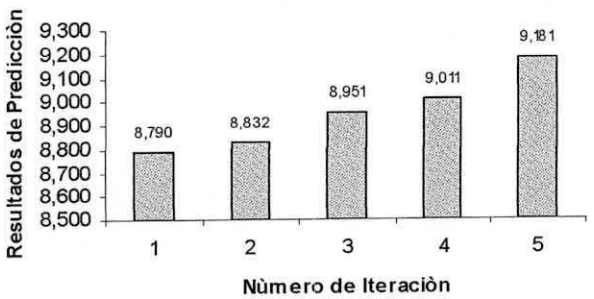
Método de Imputación por Regresión

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

Imputaciones sucesivas para $X_{41,3}$=9.189

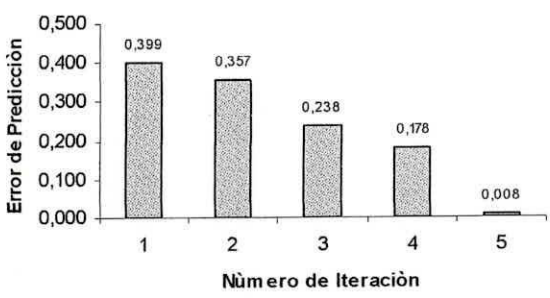
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	8.790	0,399
2	8.832	0,357
3	8.951	0,238
4	9.011	0,178
5	9.181	0,008

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	8.953
Error Estándar	0.069

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.236
Error Estándar	0.069

Elaborado por: G. Cuenca

El Cuadro 4.33, nos muestra que el primer resultado de predicción es 8.790 ± 0.399, y el último es 9.181 ± 0.399, donde la media de los resultados de predicción es 8.953 ± 0.069 y la media del error de predicción es 0.253 ± 0.069.

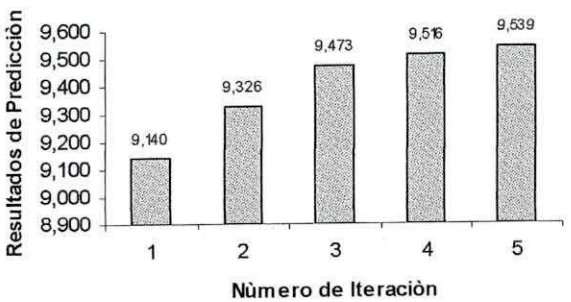
CUADRO 4.34

Efectos de la Imputación en el análisis de datos multivariados
Variables aleatorias dependientes con distribución Normal (10,1)
Método de Imputación por Regresión
Tamaño de muestra n=50 y 5% de datos faltantes en la matriz

Imputaciones sucesivas para $X_{46,3}$=9.549

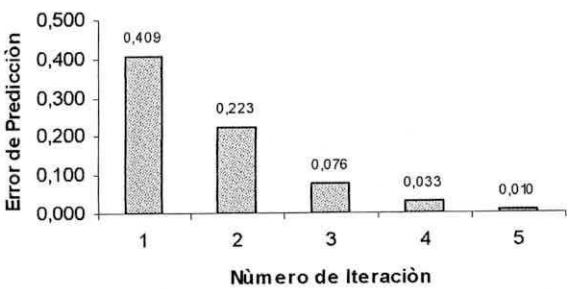
Iteración	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
1	9.140	0,409
2	9.326	0,223
3	9.473	0,076
4	9.516	0,033
5	9.539	0,010

Distribución del Resultado de Predicción



Estimadores	Resultado de Predicción
Número de Iteración	5
Media	9.399
Error Estándar	0.075

Distribución del Error de Predicción



Estimadores	Error de Predicción
Número de Iteración	5
Media	0.150
Error Estándar	0.075

Elaborado por: G. Cuenca

El Cuadro 4.34, nos muestra que el primer resultado de predicción es 9.140 ± 0.075, y el último es 9.539 ± 0.075, donde la media de los datos de predicción es 9.399 ± 0.075 y la media del error de predicción es 0.150

± 0.075. Todos los resultados de predicción de los cuadros anteriores, tienden al dato observado.

CUADRO 4.35

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

Método de Imputación por la Media y Regresión

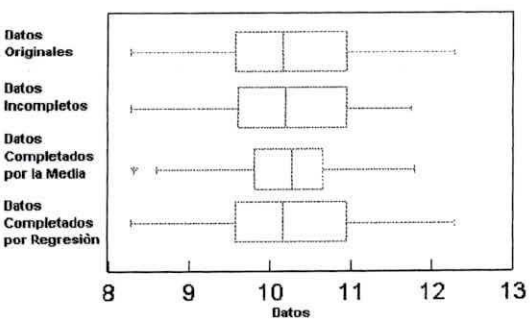
Tamaño de muestra n=50 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_3"

Estimadores

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		50	37	50	50
Media		10,214	10,194	10,194	10,216
Mediana		10,133	10,164	10,194	10,138
Moda		8,240	8,240	10,190	8,240
Varianza		0,828	0,812	0,597	0,823
Desviación Estándar		0,910	0,901	0,773	0,907
Error Estándar		0,129	0,148	0,109	0,128
Coeficiente de Asimetría		0,173	-0,012	-0,013	0,175
Curtosis		-0,639	-0,759	0,074	-0,638
Rango		4,020	3,470	3,470	4,010
Mínimo		8,240	8,240	8,240	8,240
Máximo		12,260	11,710	11,710	12,250
Percentiles	25	9,549	9,564	9,699	9,544
	50	10,133	10,164	10,194	10,138
	75	10,907	10,915	10,632	10,908

Diagrama de Cajas



Elaborado por: G. Cuenca

Al realizar la imputación por la media y regresión se obtuvieron los siguientes resultados (Ver Cuadro 4.35):

El valor de la media de los "datos completados" por *la media* disminuye. comparándolo con los "datos originales" y completados por *regresión.*

El valor de la varianza de los "datos completados" por la *media* disminuye de 0.828 a 0.597, mientras que en los "datos completados" por regresión este valor se incrementa a 0.823, comparándolo con el valor anterior y es muy cercano al valor de la varianza de los datos originales.

El vector de medias con trece datos completados por la media en X_3 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 9.997 \\ 10.012 \\ 10.194 \\ 10.137 \\ 10.188 \end{pmatrix}$$

Mientras que el vector de medias con trece datos completados por la regresión en X_3 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 9.997 \\ 10.012 \\ 10.216 \\ 10.137 \\ 10.188 \end{pmatrix}$$

El efecto que causa en la *matriz de varianzas y covarianzas* y *matriz de correlaciones*, el completar 5% de datos faltantes en una matriz de tamaño 50, por medio de la *imputación por media y regresión*, se presenta en el Cuadro 4.36.

CUADRO 4.36

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Normal (10,1)

Método de Imputación por la Media y Regresión

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

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	0.758				
X_2	0.387	0.953			
X_3	0.439	0.465	0.828		
X_4	0.135	0.439	0.396	0.517	
X_5	0.317	0.483	0.363	0.327	0.668

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.455	1.000			
X_3	0.554	0.524	1.000		
X_4	0.215	0.625	0.606	1.000	
X_5	0.445	0.606	0.488	0.556	1.000

Matriz de Varianzas y Covarianzas 26% Datos Completados por Media en "*Variable* X_3"

	X_1	X_2	X_3	X_4	X_5
X_1	0.758				
X_2	0.387	0.953			
X_3	0.262	0.304	0.597		
X_4	0.135	0.439	0.302	0.517	
X_5	0.317	0.483	0.327	0.327	0.668

Matriz de Correlaciones 26% Datos Completados por Media en "*Variable* X_3"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.455	1.000			
X_3	0.390	0.403	1.000		
X_4	0.215	0.625	0.544	1.000	
X_5	0.445	0.606	0.518	0.556	1.000

Matriz de Varianzas y Covarianzas 26% Datos Completados por Regresión en "*Variable* X_3"

	X_1	X_2	X_3	X_4	X_5
X_1	0.758				
X_2	0.387	0.953			
X_3	0.438	0.466	0.823		
X_4	0.135	0.439	0.396	0.517	
X_5	0.317	0.483	0.365	0.327	0.668

Matriz de Correlaciones 26% Datos Completados por Regresiòn en "*Variable* X_3"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.455	1.000			
X_3	0.554	0.526	1.000		
X_4	0.215	0.625	0.607	1.000	
X_5	0.445	0.606	0.493	0.556	1.000

Elaborado por: G. Cuenca

Analizando el Cuadro 4.36 se puede notar que la covarianza entre X_2 y X_3 disminuye de 0.465 a 0.304 en la matriz con 26% de "datos completados" por la media en la variable X_3, así como también la covarianza entre X_3 y X_4 disminuye 0.396 a 0.302.

En la matriz de varianzas y covarianzas de los datos completados por regresión, el valor de las covarianzas de variable X_3 con las demás variables se incrementa, comparándolo con la matriz de varianzas y covarianzas de los datos completados por la media.

Por otro lado, analizando el efecto que causa en la matriz de correlaciones, podemos apreciar en el Cuadro 4.36 que la mayor correlación se da entre las variables X_2 y X_4, es decir 0.625, seguida por 0.606 entre las variables X_2 y X_5. En la matriz de correlaciones con 26% de datos completados por la media, la correlación entre X_1 y X_3 disminuye de 0.554 a 0.390, mientras que en la matriz de datos completados por regresión, este valor es igual al de la matriz de datos originales es decir 0.554.

4.3.2 Distribución Poisson: *Cincuenta datos faltantes* en una sola variable (10% de la matriz), tamaño de muestra n=100

Se tiene una matriz de datos cuyas columnas son muestras tomadas de cinco poblaciones todas ellas Poisson, dependientes e idénticamente distribuidas, con parámetro $\lambda = 10$, $\mathbf{X} \in \mathrm{M}_{100\mathrm{x}5}$, *i= 1,2,....100* y *j= 1,2,3,4,5* y se supone que tiene el 10% de datos faltantes, es decir cincuenta datos, los que recayeron en la variable X_4 y son: el $X_{1,1}$=11, $X_{2,1}$=15, $X_{4,1}$=15, $X_{5,1}$=9, $X_{8,1}$=8, $X_{9,1}$=13, $X_{10,1}$=8, $X_{12,1}$=11, $X_{15,1}$=13, $X_{16,1}$=10, $X_{18,1}$=9, $X_{22,1}$=10, $X_{23,1}$=12, $X_{24,1}$=12, $X_{25,1}$=10, $X_{26,1}$=10, $X_{27,1}$=19, $X_{28,1}$=9, $X_{30,1}$=8, $X_{33,1}$=11, $X_{34,1}$=10, $X_{36,1}$=10, $X_{39,1}$=9, $X_{41,1}$=8, $X_{44,1}$=9, $X_{45,1}$=8,

$X_{47,1}$=11, $X_{49,1}$=10, $X_{51,1}$=9, $X_{54,1}$=6, $X_{55,1}$=12, $X_{58,1}$=8, $X_{60,1}$=8, $X_{62,1}$=10, $X_{64,1}$=12, $X_{67,1}$=9, $X_{69,1}$=9, $X_{70,1}$=12, $X_{72,1}$=10, $X_{75,1}$=8, $X_{79,1}$=4, $X_{82,1}$=12, $X_{85,1}$=14, $X_{88,1}$=15, $X_{90,1}$=9, $X_{93,1}$=13, $X_{95,1}$=11, $X_{97,1}$=11, $X_{99,1}$=11 y $X_{100,1}$=8. Nótese que el 10% de datos faltantes en la matriz, constituye 50% de datos faltantes en la columna que corresponde a X_4.

Tabla 4.26
Efectos de la Imputación en el análisis de datos multivariados
Matriz de Datos de variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Tamaño de muestra n=100

X_1	X_2	X_3	X_4	X_5
11	10	9	**11**	9
15	16	14	**15**	14
9	7	8	**6**	6
11	12	13	**15**	13
9	9	8	**9**	9
10	11	11	**12**	10
10	12	11	**12**	11
9	9	9	**8**	9
13	11	12	**13**	12
9	8	9	**8**	8
11	13	13	**12**	11
11	11	12	**11**	10
9	8	10	**9**	10
10	12	11	**11**	9
13	13	14	**13**	12
10	9	11	**10**	12
8	8	7	**7**	8
8	7	7	**9**	9
11	13	12	**11**	12
14	12	13	**14**	11
8	9	10	**10**	8
12	11	11	**10**	12
11	10	13	**12**	11
13	11	11	**12**	13
9	9	11	**10**	11
9	10	11	**10**	11
8	8	8	**9**	10
10	11	12	**9**	8
11	13	11	**12**	13
7	9	9	**8**	7
9	10	11	**10**	11
10	9	8	**9**	8
10	11	9	**11**	9
11	10	9	**10**	11

Elaborado por: G. Cuenca

Continúa...

Viene...

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Tamaño de muestra n=100

X_1	X_2	X_3	X_4	X_5
10	13	13	11	13
11	9	8	10	8
12	10	9	10	8
10	10	12	13	12
10	12	12	9	11
10	14	11	12	14
9	8	10	8	11
10	12	11	11	9
9	7	8	8	9
11	8	11	9	8
8	11	10	8	11
8	9	8	10	10
11	12	10	11	11
12	11	10	13	12
12	12	13	10	10
6	8	7	7	8
9	10	10	9	9
10	12	9	10	9
7	8	6	7	6
7	6	9	6	8
11	10	12	12	14
10	12	11	10	10
9	8	9	9	7
10	9	10	8	10
10	14	10	14	14
8	11	9	8	10
5	5	8	7	8
10	11	12	10	11
8	8	9	10	9
18	10	11	12	10
10	9	12	13	9
12	13	12	11	11
9	12	11	9	8
14	8	14	10	9
8	11	11	9	12
11	10	8	12	11
11	9	8	11	11
11	8	9	10	9
10	12	13	11	10
11	9	10	11	12
11	8	11	8	8
10	11	12	10	11
11	12	13	10	10
9	9	10	9	8
4	5	5	4	4
8	10	5	8	11
9	11	12	9	8
9	13	11	12	10
12	12	9	11	10

Elaborado por: G. Cuenca

Viene...

Efectos de la Imputación en el análisis de datos multivariados
Matriz de Datos de variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Tamaño de muestra n=100

X_1	X_2	X_3	X_4	X_5
9	6	7	8	6
10	11	14	**14**	12
13	11	13	14	12
8	9	7	7	8
13	15	13	**15**	14
15	14	11	12	11
9	8	12	**9**	10
8	9	9	10	11
10	11	10	13	12
9	8	12	**13**	10
12	10	11	9	11
11	12	10	**11**	11
13	11	9	10	9
13	11	13	**11**	13
11	10	13	14	13
10	10	11	**11**	12
8	10	9	**8**	9

Elaborado por: G. Cuenca

El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 10.120 \\ 10.200 \\ 10.360 \\ 10.270 \\ 10.140 \end{pmatrix}$$

Método de Eliminación por Filas

Debido a que los datos faltantes recayeron en la variable X_4, se procede a prescindir de las filas que tienen estos valores "faltantes", donde la matriz de datos resultante con filas eliminadas se muestra en la Tabla 4.27.

Tabla 4.27

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

Matriz de datos con cincuenta filas eliminadas

X_1	X_2	X_3	X_4	X_5
9	7	8	6	6
10	11	11	12	10
10	12	11	12	11
11	13	13	12	11
9	8	10	9	10
10	12	11	11	9
8	8	7	7	8
11	13	12	11	12
14	12	13	14	11
8	9	10	10	8
11	13	11	12	13
9	10	11	10	11
10	9	8	9	8
10	13	13	11	13
12	10	9	10	8
10	10	12	13	12
10	14	11	12	14
10	12	11	11	9
9	7	8	8	9
8	9	8	10	10
12	11	10	13	12
6	8	7	7	8
10	12	9	10	9
7	8	6	7	6
10	12	11	10	10
9	8	9	9	7
10	14	10	14	14
5	5	8	7	8
8	8	9	10	9
10	9	12	13	9
12	13	12	11	11
14	8	14	10	9
11	9	8	11	11
10	12	13	11	10
11	9	10	11	12
10	11	12	10	11
11	12	13	10	10
9	9	10	9	8
8	10	5	8	11
9	11	12	9	8
12	12	9	11	10
9	6	7	8	6
13	11	13	14	12
8	9	7	7	8
15	14	11	12	11
8	9	9	10	11
10	11	10	13	12
12	10	11	9	11
13	11	9	10	9
11	10	13	14	13

El vector de medias para las cincuenta filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 10.040 \\ 10.280 \\ 10.140 \\ 10.360 \\ 9.980 \end{pmatrix}$$

Se analiza el efecto que causa en la *matriz de varianzas y covarianzas*, y *matriz de correlaciones*, la eliminación de cincuenta filas, con un tamaño de muestra *n*=100.(Ver Cuadro 4.37)

CUADRO 4.37

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Método de Eliminación por Filas

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	4.349				
X_2	2.400	4.364			
X_3	2.421	2.493	4.091		
X_4	2.927	2.986	2.851	4.563	
X_5	2.023	2.679	2.343	3.113	3.920

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.551	1.000			
X_3	0.574	0.590	1.000		
X_4	0.657	0.669	0.660	1.000	
X_5	0.490	0.648	0.585	0.736	1.000

Matriz de Varianzas y Covarianzas (Cincuenta Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	3.835				
X_2	2.356	4.655			
X_3	2.443	2.572	4.490		
X_4	2.455	2.897	2.928	4.194	
X_5	1.613	2.863	2.146	3.069	3.979

Matriz de Correlaciones (Cincuenta Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.558	1.000			
X_3	0.589	0.563	1.000		
X_4	0.612	0.656	0.675	1.000	
X_5	0.413	0.665	0.508	0.751	1.000

Elaborado por: G. Cuenca

Se puede apreciar que la mayor covarianza en la matriz de datos originales se da entre las variables X_4 y X_5 es decir 3.113; mientras que en la matriz con cincuenta filas eliminadas este valor aumenta a 3.069.

En la matriz de correlaciones de datos originales, la mayor correlación se da entre las variables X_4 y X_5, es decir 0.736, cuyo valor se incrementa a 0.751 en la matriz de correlaciones con cincuenta filas eliminadas y por lo tanto se convierte en la mayor correlación.

CUADRO 4.38

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Método de Eliminación por Filas

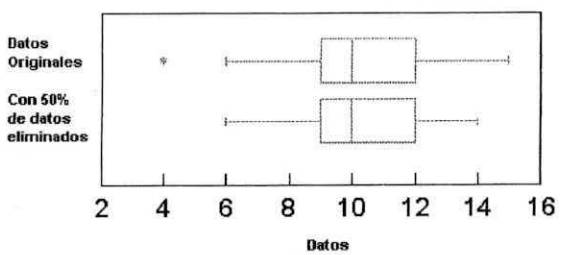
Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_4"

Estimadores

Estimadores		Datos Originales	Con el 50% de datos eliminadas en X_4
n		100	50
Media		10,270	10,360
Mediana		10,000	10,000
Moda		10,000	10,000
Varianza		4,563	4,194
Desviación Estándar		2,136	2,048
Error Estándar		0,214	0,290
Coeficiente de Asimetría		0,039	-0,080
Curtosis		0,058	-0,500
Rango		11,000	8,000
Mínimo		4,000	6,000
Máximo		15,000	14,000
Percentiles	25	9,000	9,000
	50	10,000	10,000
	75	12,000	12,000

Diagrama de Cajas



Elaborado por: G. Cuenca

En el Cuadro 4.38, podemos apreciar que con el 50% de datos eliminados en la cuarta columna de la matriz de datos (Variable X_4), el valor de la media aumentó de 10.270 a 10.360. La varianza de la variable X_4, con 50% de datos eliminados disminuyó de 4.536 a 4.194.

Método de Imputación por la Media y Regresión

A continuación se aplica el *método de imputación por media y regresión* a la misma matriz de datos utilizada en el método de eliminación por filas, es decir se completan datos en la variable X_4 que presenta cincuenta valores faltantes que son: el $X_{1,1}$=11, $X_{2,1}$=15, $X_{4,1}$=15, $X_{5,1}$=9, $X_{8,1}$=8, $X_{9,1}$=13, $X_{10,1}$=8, $X_{12,1}$=11, $X_{15,1}$=13, $X_{16,1}$=10, $X_{18,1}$=9, $X_{22,1}$=10, $X_{23,1}$=12, $X_{24,1}$=12, $X_{25,1}$=10, $X_{26,1}$=10, $X_{27,1}$=19, $X_{28,1}$=9, $X_{30,1}$=8, $X_{33,1}$=11, $X_{34,1}$=10, $X_{36,1}$=10, $X_{39,1}$=9, $X_{41,1}$=8, $X_{44,1}$=9, $X_{45,1}$=8, $X_{47,1}$=11, $X_{49,1}$=10, $X_{51,1}$=9, $X_{54,1}$=6, $X_{55,1}$=12, $X_{58,1}$=8, $X_{60,1}$=8, $X_{62,1}$=10, $X_{64,1}$=12, $X_{67,1}$=9, $X_{69,1}$=9, $X_{70,1}$=12, $X_{72,1}$=10, $X_{75,1}$=8, $X_{79,1}$=4, $X_{82,1}$=12, $X_{85,1}$=14, $X_{88,1}$=15, $X_{90,1}$=9, $X_{93,1}$=13, $X_{95,1}$=11, $X_{97,1}$=11, $X_{99,1}$=11 y $X_{100,1}$=8.

Por medio del *Método de Imputación por Media*, se procede a calcular la media aritmética de la variable X_4 con los cincuenta datos faltantes, cuyo valor es 10.360 y se reemplaza en los datos faltantes descritos anteriormente. La matriz de datos resultante con cincuenta valores completados por *imputación por la media* y *regresión* en la variable X_4 se muestra en la Tabla 4.28 y 4.29 respectivamente.

Tabla 4.28

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Método de Imputación por la Media

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
11	10	9	10.360	9
15	16	14	10.360	14
9	7	8	6	6
11	12	13	10.360	13
9	9	8	10.360	9
10	11	11	12	10
10	12	11	12	11
9	9	9	10.360	9
13	11	12	10.360	12
9	8	9	10.360	8
11	13	13	12	11
11	11	12	10.360	10
9	8	10	9	10
10	12	11	11	9
13	13	14	10.360	12
10	9	11	10.360	12
8	8	7	7	8
8	7	7	10.360	9
11	13	12	11	12
14	12	13	14	11
8	9	10	10	8
12	11	11	10.360	12
11	10	13	10.360	11
13	11	11	10.360	13
9	9	11	10.360	11
9	10	11	10.360	11
8	8	8	10.360	10
10	11	12	10.360	8
11	13	11	12	13
7	9	9	10.360	7
9	10	11	10	11
10	9	8	9	8
10	11	9	10.360	9
11	10	9	10.360	11
10	13	13	11	13
11	9	8	10.360	8
12	10	9	10	8
10	10	12	13	12
10	12	12	10.360	11
10	14	11	12	14
9	8	10	10.360	11
10	12	11	11	9
9	7	8	8	9
11	8	11	10.360	8
8	11	10	10.360	11
8	9	8	10	10
11	12	10	10.360	11
12	11	10	13	12
12	12	13	10.360	10

Continúa...

Viene...

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Método de Imputación por la Media

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
6	8	7	7	8
9	10	10	10.360	9
10	12	9	10	9
7	8	6	7	6
7	6	9	10.360	8
11	10	12	10.360	14
10	12	11	10	10
9	8	9	9	7
10	9	10	10.360	10
10	14	10	14	14
8	11	9	10.360	10
5	5	8	7	8
10	11	12	10.360	11
8	8	9	10	9
18	10	11	10.360	10
10	9	12	13	9
12	13	12	11	11
9	12	11	10.360	8
14	8	14	10	9
8	11	11	10.360	12
11	10	8	10.360	11
11	9	8	11	11
11	8	9	10.360	9
10	12	13	11	10
11	9	10	11	12
11	8	11	10.360	8
10	11	12	10	11
11	12	13	10	10
9	9	10	9	8
4	5	5	10.360	4
8	10	5	8	11
9	11	12	9	8
9	13	11	10.360	10
12	12	9	11	10
9	6	7	8	6
10	11	14	10.360	12
13	11	13	14	12
8	9	7	7	8
13	15	13	10.360	14
15	14	11	12	11
9	8	12	10.360	10
8	9	9	10	11
10	11	10	13	12
9	8	12	10.360	10
12	10	11	9	11
11	12	10	10.360	11
13	11	9	10	9
13	11	13	10.360	13
11	10	13	14	13
10	10	11	10.360	12
8	10	9	10.360	9

Tabla 4.29

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Método de Imputación por Regresión

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
11	10	9	10,979	9
15	16	14	15,064	14
9	7	8	6	6
11	12	13	14,987	13
9	9	8	9,057	9
10	11	11	12	10
10	12	11	12	11
9	9	9	8,514	9
13	11	12	12,995	12
9	8	9	8,091	8
11	13	13	12	11
11	11	12	11,015	10
9	8	10	9	10
10	12	11	11	9
13	13	14	13,048	12
10	9	11	10,031	12
8	8	7	7	8
8	7	7	8,982	9
11	13	12	11	12
14	12	13	14	11
8	9	10	10	8
12	11	11	10,018	12
11	10	13	11,924	11
13	11	11	12,081	13
9	9	11	10,005	11
9	10	11	10,012	11
8	8	8	9,071	10
10	11	12	9,100	8
11	13	11	12	13
7	9	9	8,005	7
9	10	11	10	11
10	9	8	9	8
10	11	9	10,985	9
11	10	9	10,972	11
10	13	13	11	13
11	9	8	9,901	8
12	10	9	10	8
10	10	12	13	12
10	12	12	9,172	11
10	14	11	12	14
9	8	10	8,051	11
10	12	11	11	9
9	7	8	8	9
11	8	11	9,053	8
8	11	10	8,003	11
8	9	8	10	10
11	12	10	11,072	11
12	11	10	13	12

Continúa...

Sigue...

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Método de Imputación por Regresión

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5
12	12	13	**10,030**	10
6	8	7	7	8
9	10	10	**9,022**	9
10	12	9	10	9
7	8	6	7	6
7	6	9	**5,987**	8
11	10	12	**12,101**	14
10	12	11	10	10
9	8	9	9	7
10	9	10	**7,983**	10
10	14	10	14	14
8	11	9	**8,003**	10
5	5	8	7	8
10	11	12	**10,002**	11
8	8	9	10	9
18	10	11	**11,978**	10
10	9	12	13	9
12	13	12	11	11
9	12	11	**9,062**	8
14	8	14	10	9
8	11	11	**9,051**	12
11	10	8	**11,971**	11
11	9	8	11	11
11	8	9	**10,101**	9
10	12	13	11	10
11	9	10	11	12
11	8	11	**8,106**	8
10	11	12	10	11
11	12	13	10	10
9	9	10	9	8
4	5	5	**4,031**	4
8	10	5	8	11
9	11	12	9	8
9	13	11	**11,931**	10
12	12	9	11	10
9	6	7	8	6
10	11	14	**13,920**	12
13	11	13	14	12
8	9	7	7	8
13	15	13	**14,933**	14
15	14	11	12	11
9	8	12	**9,010**	10
8	9	9	10	11
10	11	10	13	12
9	8	12	**12,915**	10
12	10	11	9	11
11	12	10	**10,993**	11
13	11	9	10	9
13	11	13	**11,061**	13
11	10	13	14	13
10	10	11	**11,076**	12
8	10	9	**8,003**	9

Tabla 4.30

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Comparación de los Métodos de Imputación

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

50% de datos completados en X_3 por la Media

Dato Observado	Resultado de Imputación por Media	Error |Dato Observado – Resultado de Imputación por Media|
11	10.360	0,64
15	10.360	4,64
15	10.360	4,64
9	10.360	1,36
8	10.360	2,36
13	10.360	2,64
8	10.360	2,36
11	10.360	0,64
13	10.360	2,64
10	10.360	0,36
9	10.360	1,36
10	10.360	0,36
12	10.360	1,64
12	10.360	1,64
10	10.360	0,36
10	10.360	0,36
9	10.360	1,36
9	10.360	1,36
8	10.360	2,36
11	10.360	0,64
10	10.360	0,36
10	10.360	0,36
9	10.360	1,36
8	10.360	2,36
9	10.360	1,36
8	10.360	2,36
11	10.360	0,64
10	10.360	0,36
9	10.360	1,36
6	10.360	4,36
12	10.360	1,64
8	10.360	2,36
8	10.360	2,36
10	10.360	0,36
12	10.360	1,64
9	10.360	1,36
9	10.360	1,36
12	10.360	1,64
10	10.360	0,36
8	10.360	2,36
4	10.360	6,36
12	10.360	1,64
14	10.360	3,64
15	10.360	4,64
9	10.360	1,36
13	10.360	2,64
11	10.360	0,64
11	10.360	0,64
11	10.360	0,64
8	10.360	2,36

Viene...

Variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Comparación de los Métodos de Imputación

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

50% de datos completados en X_3 por Regresión

Dato Observado	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
11	10.979	0,021
15	15.064	0,064
15	14.987	0,013
9	9.057	0,057
8	8.514	0,514
13	12.995	0,005
8	8.091	0,091
11	11.015	0,015
13	13.048	0,048
10	10.031	0,031
9	8.982	0,018
10	10.018	0,018
12	11.924	0,076
12	12.081	0,081
10	10.005	0,005
10	10.012	0,012
9	9.071	0,071
9	9.100	0,100
8	8.005	0,005
11	10.985	0,015
10	10.972	0,972
10	8.901	1,099
9	9.172	0,172
8	8.051	0,051
9	9.053	0,053
8	8.003	0,003
11	11.072	0,072
10	10.030	0,030
9	9.022	0,022
6	5.987	0,013
12	12.101	0,101
8	7.983	0,017
8	8.003	0,003
10	10.002	0,002
12	11.978	0,022
9	9.062	0,062
9	9.051	0,051
12	11.971	0,029
10	10.101	0,101
8	8.106	0,106
4	4.031	0,031
12	11.931	0,069
14	13.920	0,080
15	14.933	0,067
9	9.010	0,010
13	12.915	0,085
11	10.993	0,007
11	11.061	0,061
11	11.076	0,076
8	8.003	0,003

Se puede notar, por medio de la Tabla 4.30 que la diferencia en valor absoluto entre el valor observado de cada variable, es menor en el *Método de Imputación por Regresión.*

CUADRO 4.39

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Método de Imputación por la Media y Regresión

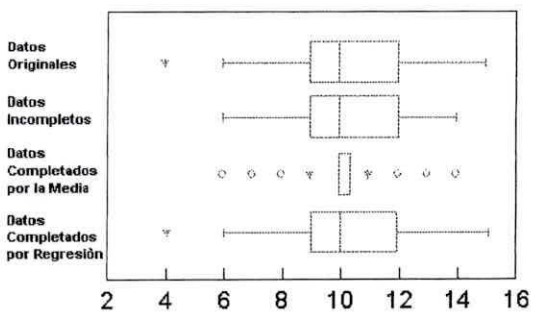
Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_4"

Estimadores

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		100	50	100	100
Media		10,270	10,360	10,360	10,295
Mediana		10,000	10,000	10,360	10,004
Moda		10,000	10,000	10,360	10,000
Varianza		4,563	4,194	2,076	4,510
Desviación Estándar		2,136	2,048	1,441	2,124
Error Estándar		0,214	0,290	0,144	0,212
Coeficiente de Asimetría		0,039	-0,080	-0,111	0,016
Curtosis		0,058	-0,500	2,022	0,087
Rango		11,000	8,000	8,000	11,030
Mínimo		4,000	6,000	6,000	4,030
Máximo		15,00	14,000	14,000	15,060
Percentiles	25	9,000	9,000	10,000	9,000
	50	10,000	10,000	10,360	10,004
	75	12,000	12,000	10,360	11,961

Diagrama de Cajas



Elaborado por: G. Cuenca

Al realizar la imputación por la media y regresión se obtuvieron los siguientes resultados (Ver Cuadro 4.39):

El valor de la media de los "datos completados" por *la media* aumenta, comparándolo con los "datos originales" y completados por *regresión.*

El valor de la varianza de los "datos completados" por la *media* disminuye de 4.563 a 2.076, mientras que en los datos completados por regresión este valor se incrementa a 4.510, comparándolo con el valor anterior y es muy cercano al valor de la varianza de los "datos originales".

El vector de medias con cincuenta datos completados por la media en X_4 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 10.120 \\ 10.200 \\ 10.360 \\ 10.360 \\ 10.140 \end{pmatrix}$$

Mientras que el vector de medias con cincuenta datos completados por la regresión en X_4 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \end{pmatrix} = \begin{pmatrix} 10.120 \\ 10.200 \\ 10.360 \\ 10.295 \\ 10.140 \end{pmatrix}$$

El efecto que causa en la *matriz de varianzas* y *covarianzas* y *matriz de correlaciones*, el completar 10% de datos faltantes en una matriz de tamaño 100, por medio de la imputación por media y regresión, se presenta en el Cuadro 4.40.

CUADRO 4.40

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Poisson $\lambda = 10$

Método de Imputación por la Media y Regresión

Tamaño de muestra n=100 y 10% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	4.349				
X_2	2.400	4.364			
X_3	2.421	2.493	4.091		
X_4	2.927	2.986	2.851	4.563	
X_5	2.023	2.679	2.343	3.113	3.920

Matriz de Correlaciones (Datos Originales)

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.551	1.000			
X_3	0.574	0.590	1.000		
X_4	0.657	0.669	0.660	1.000	
X_5	0.490	0.648	0.585	0.736	1.000

Matriz de Varianzas y Covarianzas 50% Datos Completados por Media en "*Variable* X_4"

	X_1	X_2	X_3	X_4	X_5
X_1	4.349				
X_2	2.400	4.364			
X_3	2.421	2.493	4.091		
X_4	1.215	1.434	1.449	2.076	
X_5	2.023	2.679	2.343	1.519	3.920

Matriz de Correlaciones 50% Datos Completados por Media en "*Variable* X_4"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.551	1.000			
X_3	0.574	0.590	1.000		
X_4	0.404	0.476	0.497	1.000	
X_5	0.490	0.648	0.585	0.532	1.000

Matriz de Varianzas y Covarianzas 50% Datos Completados por Regresión en "*Variable* X_4"

	X_1	X_2	X_3	X_4	X_5
X_1	4.349				
X_2	2.400	4.364			
X_3	2.421	2.493	4.091		
X_4	2.931	2.976	2.834	4.510	
X_5	2.023	2.679	2.343	3.118	3.920

Matriz de Correlaciones 50% Datos Completados por Regresiòn en "*Variable* X_4"

	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.551	1.000			
X_3	0.574	0.590	1.000		
X_4	0.662	0.671	0.660	1.000	
X_5	0.490	0.648	0.585	0.742	1.000

Elaborado por: G. Cuenca

La covarianza entre X_4 y X_5 disminuye de 3.113 a 1.159 en la matriz con 50% de "datos completados" por la media en la variable X_4, así como también disminuye la covarianza entre X_4 con las otras variables.

En la matriz de varianzas y covarianzas de los datos completados por regresión, el valor de las covarianzas de variable X_4 con las demás variables se incrementa, comparándolo con la matriz de varianzas y covarianzas de los "datos completados" por *la media.*

Por otro lado, analizando el efecto que causa en la matriz de correlaciones, podemos apreciar en le Cuadro 4.40 que también los únicos valores que cambian son los de la correlación de X_4 con las demás variables, puesto que a esta variable se le completó datos por medio de los métodos de imputación; donde la mayor correlación se da entre las variables X_4 y X_5, es decir 0.736, seguida por 0.669 entre las variables X_2 y X_4. En la matriz de correlaciones con 50% de datos completados por la media, la correlación entre X_4 y X_5 disminuye de 0.736 a 0.532, mientras que en la matriz de datos completados por regresión, este valor es 0.742.

4.3.3 Distribución Exponencial: *Cincuenta datos faltantes*: Veinticinco en X_3 y veinticinco en X_8 (10% de la matriz), tamaño de muestra n=100

Se tiene una matriz de datos cuyas columnas son muestras tomadas de diez poblaciones todas ellas Exponencial, dependientes e idénticamente distribuidas, con parámetro $\beta = 4$, $\mathbf{X} \in \mathrm{M}_{100\mathrm{x}10}$, *i= 1,2,....100* y *j= 1,2,3,...,10* y se supone que tiene el 5% de datos faltantes, es decir

cincuenta datos, los que recayeron en las variables X_3 y X_8 y son: el $X_{3,3}$=2.851, $X_{9,3}$=1.414, $X_{15,3}$=1.069, $X_{18,3}$=6.462, $X_{21,3}$=3.914, $X_{24,3}$=1.131, $X_{31,3}$=6.562, $X_{33,3}$=2.254, $X_{39,3}$=1.689, $X_{42,3}$=1.432, $X_{43,3}$=3.693, $X_{47,3}$=3.960, $X_{48,3}$=3.420, $X_{52,3}$=2.683, $X_{55,3}$=6.730, $X_{58,3}$=0.860, $X_{59,3}$=6.406, $X_{67,3}$=3.578, $X_{69,3}$=5.157, $X_{71,3}$=4.083, $X_{74,3}$=2.061, $X_{79,3}$=1.148, $X_{81,3}$=3.359, $X_{84,3}$=1.913, $X_{86,3}$=1.351, $X_{6,8}$=2.390, $X_{12,8}$=1.060, $X_{17,8}$=1.383, $X_{23,8}$=1.219, $X_{30,8}$=2.582, $X_{34,8}$=5.997, $X_{37,8}$=3.952, $X_{41,8}$=19.664, $X_{46,8}$=5.859, $X_{50,8}$=5.255, $X_{53,8}$=9.518, $X_{60,8}$=2.947, $X_{61,8}$=2.566, $X_{62,8}$=0.929, $X_{63,8}$=4.580, $X_{75,8}$=2.080, $X_{77,8}$=3.767, $X_{87,8}$=4.930, $X_{88,8}$=6.314, $X_{92,8}$=0.704, $X_{93,8}$=5.413, $X_{97,8}$=3.183, $X_{98,8}$=4.859, $X_{99,8}$=4.800 y $X_{100,8}$=5.525.

Nótese que el 5% de datos faltantes en la matriz, constituye 25% de datos faltantes en la columna que corresponde a X_3 y 25% de datos faltantes en la columna X_8 (Ver Tabla 4.31)

Tabla 4.31

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Tamaño de muestra n=100

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
6.726	6.168	3.447	4.124	4.017	4.550	5.149	4.957	6.743	3.346
1.168	1.763	0.622	2.786	4.782	3.397	3.994	1.921	1.714	2.373
3.238	4.557	2.851	3.335	0.641	10.599	10.406	11.662	0.222	10.237
0.283	0.163	0.814	2.302	1.101	2.715	0.470	0.462	0.814	2.980
3.054	1.277	3.099	1.934	0.206	1.929	0.575	1.089	0.289	1.435
3.483	3.547	3.129	5.710	3.334	3.645	5.478	2.390	4.686	3.469
0.668	1.180	3.188	2.429	4.009	3.122	2.252	1.105	4.255	2.085
2.576	2.268	3.545	1.127	2.069	3.408	3.349	3.863	2.491	3.414
4.385	1.285	1.414	1.937	1.812	2.162	2.081	4.421	4.249	4.599
1.589	1.276	2.751	0.819	2.093	2.700	2.421	2.740	2.224	2.820
0.706	1.523	4.851	1.602	4.022	1.399	1.671	2.287	4.115	1.108
1.721	3.194	1.051	3.420	1.406	3.575	1.586	1.060	1.712	3.696
1.535	1.701	1.466	1.192	2.600	3.875	2.265	1.995	1.767	3.724
3.876	1.856	1.723	1.872	2.278	1.143	1.079	2.902	1.891	2.860
0.737	2.047	1.069	2.488	1.351	1.041	2.934	2.882	1.617	1.052
2.750	5.298	2.372	5.287	5.913	4.634	4.520	3.012	4.673	3.123
1.373	1.996	3.664	1.678	3.197	1.797	2.731	1.383	2.728	1.343
3.386	1.849	6.462	5.218	6.036	2.054	6.604	2.182	1.310	2.984
4.755	3.972	1.879	3.576	2.127	2.750	1.792	1.623	2.187	3.749
2.650	2.213	1.241	2.986	2.135	1.215	1.608	1.562	1.126	1.524
5.571	3.181	3.914	5.382	3.060	3.755	1.035	4.237	5.737	5.339
1.530	2.504	2.470	2.068	1.122	0.344	3.872	1.045	3.311	1.349
4.779	4.420	3.471	4.447	0.445	4.719	3.270	1.219	4.179	3.091
2.452	4.650	1.131	2.951	4.005	0.832	2.911	2.574	2.371	1.803
2.565	2.414	0.923	2.062	5.526	2.385	1.990	2.036	2.973	2.421
1.439	3.829	1.334	1.294	1.279	2.422	2.949	2.741	1.932	2.659
3.888	1.524	3.675	4.748	7.131	7.411	7.808	1.854	5.252	5.882
1.603	1.507	4.001	2.180	1.244	1.084	2.942	1.930	2.045	1.612
2.633	1.371	1.907	2.073	1.416	1.304	2.665	3.206	1.354	1.596
2.086	1.962	1.252	1.197	1.661	1.713	2.182	2.582	2.399	2.791
2.800	1.987	6.562	1.832	6.257	1.129	6.075	7.053	1.242	6.120
7.423	6.601	6.400	3.976	3.149	1.643	7.398	7.141	4.436	6.879
3.786	6.453	2.254	6.418	6.050	5.496	3.591	6.079	1.401	3.806
1.755	6.641	1.837	5.535	3.645	5.206	3.588	5.997	3.233	1.775
0.804	2.132	5.803	3.424	2.305	3.475	7.773	7.824	2.168	4.732
1.661	1.418	2.400	3.917	4.567	1.186	1.240	3.133	1.511	1.656
4.292	4.003	3.284	4.179	3.924	4.342	4.589	3.952	1.153	4.109
4.955	2.839	4.372	3.730	3.567	3.045	3.825	5.077	3.874	2.255
4.301	1.327	1.689	2.704	3.954	2.647	4.671	2.970	1.283	2.873
2.509	1.469	3.747	3.180	7.432	4.313	7.123	4.382	7.261	4.588
1.275	9.904	1.865	1.178	6.441	3.053	1.436	19.664	0.179	1.579
4.694	3.156	1.432	7.665	6.024	4.361	4.524	2.119	6.514	2.655
0.705	3.267	3.693	0.557	2.272	2.904	1.237	2.449	1.013	2.028
2.262	4.162	3.531	1.048	1.417	1.594	3.558	1.702	1.956	1.286
3.973	3.493	1.691	3.246	2.600	4.683	3.667	4.641	3.274	4.739
1.411	1.568	0.709	1.908	2.580	1.461	2.729	5.859	2.888	0.146
2.416	1.431	3.960	1.198	1.046	2.869	6.104	3.508	4.971	6.288
3.240	1.273	3.420	1.785	3.923	4.030	2.579	4.832	3.118	4.303
1.458	2.949	2.079	3.588	1.777	3.941	1.778	1.587	1.203	1.796
4.904	5.356	5.279	5.169	10.262	5.529	10.492	5.255	5.913	10.542

Elaborado por: G. Cuenca

Continúa...

Sigue...

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Tamaño de muestra n=100

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.919	4.446	1.333	4.688	2.057	4.830	0.712	4.278	0.169	4.367
0.177	3.187	2.683	2.848	0.209	2.757	0.875	2.298	2.544	2.163
6.596	7.337	7.012	6.574	7.968	6.449	9.439	9.518	11.604	11.447
0.511	0.453	2.859	3.076	0.471	1.660	3.090	3.111	2.044	3.462
7.019	5.192	6.730	1.971	7.147	1.580	7.974	5.108	3.734	3.566
5.082	5.470	6.423	10.571	10.914	6.174	5.790	4.342	6.142	7.809
5.216	0.577	0.070	4.630	5.805	6.604	6.580	2.890	6.806	0.555
5.508	5.838	0.860	1.988	0.277	8.785	3.236	0.196	8.269	9.167
7.531	7.868	6.406	6.971	10.050	5.283	10.384	5.845	6.612	5.274
1.219	2.384	2.895	1.906	5.324	2.125	4.701	2.947	2.949	2.660
7.965	5.846	8.665	5.610	5.002	4.962	4.533	2.566	6.117	4.267
0.773	2.720	1.633	2.129	1.205	0.556	0.720	0.929	0.521	0.184
0.449	2.003	4.027	2.725	2.785	2.466	4.397	4.580	4.170	2.684
6.455	7.185	7.863	3.065	4.945	2.619	1.508	1.379	1.302	1.192
4.833	1.780	2.271	2.454	1.586	2.595	2.939	1.324	1.128	4.257
1.653	2.624	0.779	0.238	0.172	1.338	2.313	1.290	1.440	2.493
5.029	3.679	3.578	4.295	3.063	5.534	4.939	4.058	5.257	4.231
3.027	6.997	3.002	3.647	1.625	2.274	1.651	3.216	4.641	1.289
4.947	4.069	5.157	4.715	5.132	4.946	4.934	0.827	4.110	4.323
1.047	1.023	4.330	3.551	4.398	2.603	1.513	1.317	4.113	1.171
2.160	3.286	4.083	5.008	5.835	4.443	5.692	6.458	6.420	6.410
3.437	4.315	2.402	3.724	4.977	2.237	3.348	3.577	4.924	3.505
2.650	4.631	4.361	2.749	4.810	4.374	2.653	2.303	2.003	4.456
4.387	4.031	2.061	1.303	2.059	3.308	2.004	4.271	4.820	3.195
3.349	2.733	2.041	4.734	3.214	3.010	2.136	2.080	1.895	2.561
5.950	5.100	5.241	8.751	8.797	5.607	6.784	5.941	8.083	5.750
2.180	4.490	1.422	3.254	2.905	3.984	4.586	3.767	4.684	5.501
1.096	3.067	1.154	3.048	2.318	2.521	2.126	1.073	4.016	4.150
2.541	3.118	1.148	1.888	3.642	1.282	3.155	0.424	3.997	1.188
1.782	2.558	1.205	1.638	2.784	3.678	3.476	1.468	1.700	1.718
5.260	2.272	3.359	1.292	4.339	4.104	2.877	3.287	3.006	2.248
3.872	3.320	1.821	3.069	1.131	3.017	1.615	1.421	3.691	2.732
0.977	5.323	3.878	5.360	1.664	1.563	3.183	1.979	1.301	3.020
1.538	0.544	1.913	1.379	4.166	1.871	2.308	4.817	3.755	1.849
3.097	3.744	2.224	2.974	2.029	3.689	3.154	0.622	2.684	3.376
2.264	1.749	1.351	1.056	2.011	1.089	1.400	1.754	2.505	2.449
1.922	1.135	2.030	2.992	1.665	1.782	3.061	4.930	3.322	3.144
5.309	1.632	5.489	0.409	6.785	5.881	5.931	6.314	7.342	5.589
2.024	2.555	3.541	3.185	1.807	1.535	2.964	3.691	1.676	1.626
1.962	1.450	2.667	3.870	4.081	1.627	3.066	4.395	4.515	3.001
3.514	6.951	1.244	2.751	2.468	2.018	2.323	1.230	4.707	1.959
0.750	0.800	0.449	1.177	1.890	1.178	2.311	0.704	0.035	1.687
0.537	1.374	1.158	5.727	1.508	5.355	1.709	5.413	1.359	1.518
10.796	10.465	10.521	8.610	8.558	8.599	8.032	10.596	8.551	8.123
1.056	1.232	2.367	1.325	0.526	1.676	2.971	2.542	2.939	2.814
8.118	8.500	9.924	9.254	10.405	9.636	10.348	9.218	9.805	10.046
4.040	4.244	3.613	3.099	4.680	6.852	2.452	3.183	4.986	5.924
4.949	7.182	4.366	3.236	4.999	3.716	3.930	4.859	7.191	4.460
3.296	7.442	7.542	4.733	4.720	3.237	3.182	4.800	5.351	3.424
5.786	6.620	6.717	5.252	5.305	5.491	6.526	5.525	6.722	4.637

Elaborado por: G. Cuenca

El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \\ \overline{X}_7 \\ \overline{X}_8 \\ \overline{X}_9 \\ \overline{X}_{10} \end{pmatrix} = \begin{pmatrix} 3.164 \\ 3.445 \\ 3.206 \\ 3.350 \\ 3.614 \\ 3.391 \\ 3.741 \\ 3.588 \\ 3.526 \\ 3.532 \end{pmatrix}$$

Método de Eliminación por Filas

Debido a que los datos faltantes recayeron en las variables X_3 y X_8 es decir en: $X_{3,3}$=2.851, $X_{9,3}$=1.414, $X_{15,3}$=1.069, $X_{18,3}$=6.462, $X_{21,3}$=3.914, $X_{24,3}$=1.131, $X_{31,3}$=6.562, $X_{33,3}$=2.254, $X_{39,3}$=1.689, $X_{42,3}$=1.432, $X_{43,3}$=3.693, $X_{47,3}$=3.960, $X_{48,3}$=3.420, $X_{52,3}$=2.683, $X_{55,3}$=6.730, $X_{58,3}$=0.860, $X_{59,3}$=6.406, $X_{67,3}$=3.578, $X_{69,3}$=5.157, $X_{71,3}$=4.083, $X_{74,3}$=2.061, $X_{79,3}$=1.148, $X_{81,3}$=3.359, $X_{84,3}$=1.913, $X_{86,3}$=1.351, $X_{6,8}$=2.390, $X_{12,8}$=1.060, $X_{17,8}$=1.383, $X_{23,8}$=1.219, $X_{30,8}$=2.582, $X_{34,8}$=5.997, $X_{37,8}$=3.952, $X_{41,8}$=19.664, $X_{46,8}$=5.859, $X_{50,8}$=5.255, $X_{53,8}$=9.518, $X_{60,8}$=2.947, $X_{61,8}$=2.566, $X_{62,8}$=0.929, $X_{63,8}$=4.580, $X_{75,8}$=2.080, $X_{77,8}$=3.767, $X_{87,8}$=4.930, $X_{88,8}$=6.314, $X_{92,8}$=0.704, $X_{93,8}$=5.413, $X_{97,8}$=3.183, $X_{98,8}$=4.859, $X_{99,8}$=4.800 y $X_{100,8}$=5.525, se procede a prescindir de las filas que tienen estos valores "faltantes"(Ver Tabla 4.32).

Tabla 4.32

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

Matriz de datos con cincuenta filas eliminadas

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
6.726	6.168	3.447	4.124	4.017	4.550	5.149	4.957	6.743	3.346
1.168	1.763	0.622	2.786	4.782	3.397	3.994	1.921	1.714	2.373
0.283	0.163	0.814	2.302	1.101	2.715	0.470	0.462	0.814	2.980
3.054	1.277	3.099	1.934	0.206	1.929	0.575	1.089	0.289	1.435
0.668	1.180	3.188	2.429	4.009	3.122	2.252	1.105	4.255	2.085
2.576	2.268	3.545	1.127	2.069	3.408	3.349	3.863	2.491	3.414
1.589	1.276	2.751	0.819	2.093	2.700	2.421	2.740	2.224	2.820
0.706	1.523	4.851	1.602	4.022	1.399	1.671	2.287	4.115	1.108
1.535	1.701	1.466	1.192	2.600	3.875	2.265	1.995	1.767	3.724
3.876	1.856	1.723	1.872	2.278	1.143	1.079	2.902	1.891	2.860
4.755	3.972	1.879	3.576	2.127	2.750	1.792	1.623	2.187	3.749
2.650	2.213	1.241	2.986	2.135	1.215	1.608	1.562	1.126	1.524
1.530	2.504	2.470	2.068	1.122	0.344	3.872	1.045	3.311	1.349
2.565	2.414	0.923	2.062	5.526	2.385	1.990	2.036	2.973	2.421
1.439	3.829	1.334	1.294	1.279	2.422	2.949	2.741	1.932	2.659
3.888	1.524	3.675	4.748	7.131	7.411	7.808	1.854	5.252	5.882
1.603	1.507	4.001	2.180	1.244	1.084	2.942	1.930	2.045	1.612
2.633	1.371	1.907	2.073	1.416	1.304	2.665	3.206	1.354	1.596
7.423	6.601	6.400	3.976	3.149	1.643	7.398	7.141	4.436	6.879
0.804	2.132	5.803	3.424	2.305	3.475	7.773	7.824	2.168	4.732
1.661	1.418	2.400	3.917	4.567	1.186	1.240	3.133	1.511	1.656
4.955	2.839	4.372	3.730	3.567	3.045	3.825	5.077	3.874	2.255
2.509	1.469	3.747	3.180	7.432	4.313	7.123	4.382	7.261	4.588
2.262	4.162	3.531	1.048	1.417	1.594	3.558	1.702	1.956	1.286
3.973	3.493	1.691	3.246	2.600	4.683	3.667	4.641	3.274	4.739
1.458	2.949	2.079	3.588	1.777	3.941	1.778	1.587	1.203	1.796
0.919	4.446	1.333	4.688	2.057	4.830	0.712	4.278	0.169	4.367
0.511	0.453	2.859	3.076	0.471	1.660	3.090	3.111	2.044	3.462
5.082	5.470	6.423	10.571	10.914	6.174	5.790	4.342	6.142	7.809
5.216	0.577	0.070	4.630	5.805	6.604	6.580	2.890	6.806	0.555
6.455	7.185	7.863	3.065	4.945	2.619	1.508	1.379	1.302	1.192
4.833	1.780	2.271	2.454	1.586	2.595	2.939	1.324	1.128	4.257
1.653	2.624	0.779	0.238	0.172	1.338	2.313	1.290	1.440	2.493
3.027	6.997	3.002	3.647	1.625	2.274	1.651	3.216	4.641	1.289
1.047	1.023	4.330	3.551	4.398	2.603	1.513	1.317	4.113	1.171
3.437	4.315	2.402	3.724	4.977	2.237	3.348	3.577	4.924	3.505
2.650	4.631	4.361	2.749	4.810	4.374	2.653	2.303	2.003	4.456
5.950	5.100	5.241	8.751	8.797	5.607	6.784	5.941	8.083	5.750
1.096	3.067	1.154	3.048	2.318	2.521	2.126	1.073	4.016	4.150
1.782	2.558	1.205	1.638	2.784	3.678	3.476	1.468	1.700	1.718
3.872	3.320	1.821	3.069	1.131	3.017	1.615	1.421	3.691	2.732
0.977	5.323	3.878	5.360	1.664	1.563	3.183	1.979	1.301	3.020
3.097	3.744	2.224	2.974	2.029	3.689	3.154	0.622	2.684	3.376
2.024	2.555	3.541	3.185	1.807	1.535	2.964	3.691	1.676	1.626
1.962	1.450	2.667	3.870	4.081	1.627	3.066	4.395	4.515	3.001
3.514	6.951	1.244	2.751	2.468	2.018	2.323	1.230	4.707	1.959
10.796	10.465	10.521	8.610	8.558	8.599	8.032	10.596	8.551	8.123
1.056	1.232	2.367	1.325	0.526	1.676	2.971	2.542	2.939	2.814
8.118	8.500	9.924	9.254	10.405	9.636	10.348	9.218	9.805	10.046
6.726	6.168	3.447	4.124	4.017	4.550	5.149	4.957	6.743	3.346

Elaborado por: G. Cuenca

El vector de medias para las cincuenta filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \\ \overline{X}_7 \\ \overline{X}_8 \\ \overline{X}_9 \\ \overline{X}_{10} \end{pmatrix} = \begin{pmatrix} 3.082 \\ 3.270 \\ 3.158 \\ 3.353 \\ 3.366 \\ 3.161 \\ 3.450 \\ 3.059 \\ 3.346 \\ 3.222 \end{pmatrix}$$

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*, y *matriz de correlaciones*, la eliminación de cincuenta filas, con un tamaño de muestra n=100.

CUADRO 4.41

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Eliminación por Filas

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas
(Datos Originales)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	4.386									
X_2	2.700	4.854								
X_3	2.701	2.336	4.528							
X_4	2.165	2.247	2.041	3.978						
X_5	2.780	2.346	3.072	2.968	6.029					
X_6	2.252	1.997	1.489	2.269	2.240	4.084				
X_7	2.706	1.857	2.925	2.329	3.696	2.695	5.563			
X_8	1.637	2.954	2.226	1.497	2.897	2.173	3.059	7.626		
X_9	3.019	2.133	2.366	2.365	3.272	2.508	3.039	1.543	5.322	
X_{10}	2.552	2.045	2.371	2.044	2.521	3.192	3.685	2.716	2.939	5.072

Matriz de Varianzas y Covarianzas
(Cincuenta Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	5.136									
X_2	3.729	5.163								
X_3	3.001	3.035	4.813							
X_4	2.889	2.730	2.849	4.413						
X_5	3.226	2.320	3.247	4.180	6.452					
X_6	2.606	1.939	2.009	2.879	3.690	3.844				
X_7	2.925	1.975	2.881	2.848	3.717	3.004	4.956			
X_8	3.041	2.650	3.279	2.847	2.932	2.358	3.560	4.705		
X_9	3.303	2.498	2.551	3.175	4.367	2.924	3.744	3.001	5.132	
X_{10}	2.468	2.255	2.569	2.878	3.074	2.689	3.055	2.958	2.433	3.854

Elaborado por: G. Cuenca

Analizando el Cuadro 4.41, se puede apreciar que la mayor covarianza en la matriz de datos originales se da entre las variables X_5 y X_9 es decir 3.272; mientras que en la matriz con cincuenta filas eliminadas este valor aumenta a 4.367.

CUADRO 4.42

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Eliminación por Filas

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

Matriz de Correlaciones
(Datos Originales)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000									
X_2	0.585	1.000								
X_3	0.606	0.498	1.000							
X_4	0.518	0.511	0.481	1.000						
X_5	0.541	0.434	0.588	0.606	1.000					
X_6	0.532	0.448	0.346	0.563	0.451	1.000				
X_7	0.548	0.357	0.583	0.495	0.638	0.565	1.000			
X_8	0.283	0.486	0.379	0.272	0.427	0.389	0.470	1.000		
X_9	0.625	0.420	0.482	0.514	0.578	0.538	0.559	0.242	1.000	
X_{10}	0.541	0.412	0.495	0.455	0.456	0.701	0.694	0.437	0.566	1.000

Matriz de Correlaciones
(Cincuenta Filas Eliminadas)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000									
X_2	0.724	1.000								
X_3	0.604	0.609	1.000							
X_4	0.607	0.572	0.618	1.000						
X_5	0.560	0.402	0.583	0.783	1.000					
X_6	0.586	0.435	0.467	0.699	0.741	1.000				
X_7	0.580	0.390	0.590	0.609	0.657	0.688	1.000			
X_8	0.619	0.538	0.689	0.625	0.532	0.554	0.737	1.000		
X_9	0.643	0.485	0.513	0.667	0.759	0.658	0.742	0.611	1.000	
X_{10}	0.555	0.505	0.596	0.698	0.616	0.699	0.699	0.695	0.547	1.000

Elaborado por: G. Cuenca

En la matriz de correlaciones de datos originales, la mayor correlación se da entre las variables X_7 y X_{10}, es decir 0.701, cuyo valor se disminuye a 0.699 en la matriz de correlaciones con cincuenta filas eliminadas. La mayor correlación en la matriz con cincuenta filas eliminadas es entre las variables X_4 y X_5, es decir 0.783. En general, se puede decir que la correlación entre las variables, se incrementó en la matriz con 50 filas eliminadas.

CUADRO 4.43

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Eliminación por Filas

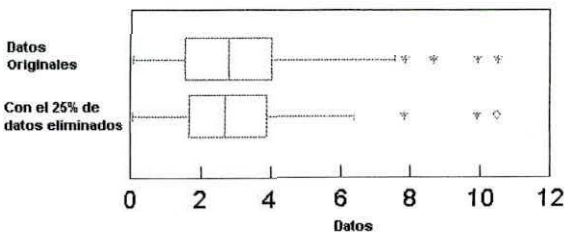
Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_3" y "*Variable* X_8"

Estimadores "*Variable* X_3"

Estimadores		Datos Originales	Con el 25% de datos eliminadas en X_3
n		100	50
Media		3,206	3,158
Mediana		2,801	2,709
Moda		0,070	3,450
Varianza		4,528	4,813
Desviación Estándar		2,128	2,194
Error Estándar		0,213	0,310
Coeficiente de Asimetría		1,194	1,559
Curtosis		1,351	2,943
Rango		10,450	10,450
Mínimo		0,070	0,070
Máximo		10,520	10,520
Percentiles	25	1,508	1,635
	50	2,801	2,709
	75	4,020	3,909

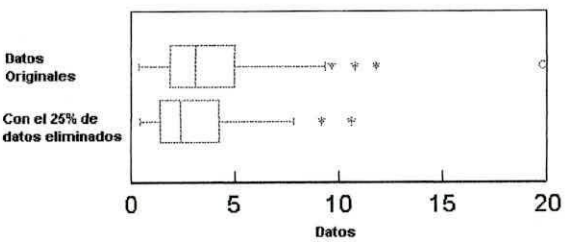
Diagrama de Cajas "*Variable* X_3"



Estimadores "*Variable* X_8"

Estimadores		Datos Originales	Con el 25% de datos eliminadas en X_8
n		100	50
Media		3,588	3,059
Mediana		2,959	2,423
Moda		0,200	4,960
Varianza		7,626	4,705
Desviación Estándar		2,762	2,169
Error Estándar		0,276	0,307
Coeficiente de Asimetría		2,576	1,619
Curtosis		11,269	2,870
Rango		19,470	10,130
Mínimo		0,200	0,460
Máximo		19,660	10,600
Percentiles	25	1,715	1,456
	50	2,959	2,423

Diagrama de Cajas "*Variable* X_8"



Elaborado por: G. Cuenca

En el Cuadro 4.43, podemos apreciar que con el 25% de datos eliminados en la tercera columna de la matriz de datos (Variable X_3), el valor de la media y la mediana disminuyó de 3.206 a 3.158 y de 2.801 a 2.709, respectivamente. La varianza de la variable X_3, con 25% de datos eliminados aumentó de 4.528 a 4.813. En la variable X_8, el valor de la media y la mediana disminuyeron su valor, así como también el valor de la varianza.

Método de Imputación por la Media y Regresión

Estos métodos se aplican a la misma matriz de datos utilizada en el método de eliminación por filas, es decir se completan datos en la variable X_3 y X_8, que presentan veinte y cinco valores faltantes cada una. A través del Método de Imputación por Media, se procede a calcular la media aritmética de la variable X_3 con los veinticinco datos faltantes, cuyo valor es 3.219, entonces reemplazamos en $X_{3,3}$, $X_{9,3}$, $X_{15,3}$, $X_{18,3}$, $X_{21,3}$, $X_{24,3}$, $X_{31,3}$, $X_{33,3}$, $X_{39,3}$, $X_{42,3}$, $X_{43,3}$, $X_{47,3}$, $X_{48,3}$, $X_{52,3}$, $X_{55,3}$, $X_{58,3}$, $X_{59,3}$, $X_{67,3}$, $X_{69,3}$, $X_{71,3}$, $X_{74,3}$, $X_{79,3}$, $X_{81,3}$, $X_{84,3}$, $X_{86,3}$, también se calcula el valor de la media de la variable X_8, 3.298, mismo que se remplaza en $X_{6,8}$, $X_{12,8}$, $X_{17,8}$, $X_{23,8}$, $X_{30,8}$, $X_{34,8}$, $X_{37,8}$, $X_{41,8}$, $X_{46,8}$, $X_{50,8}$, $X_{53,8}$, $X_{60,8}$, $X_{61,8}$, $X_{62,8}$, $X_{63,8}$, $X_{75,8}$, $X_{77,8}$, $X_{87,8}$, $X_{88,8}$, $X_{92,8}$, $X_{93,8}$, $X_{97,8}$, $X_{98,8}$, $X_{99,8}$ y en $X_{100,8}$. La matriz de datos resultante con cincuenta valores completados por *imputación por la media* y *regresión*, se muestra en la Tabla 4.33 y 4.34 respectivamente.

Tabla 4.33

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Imputación por Media

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
6.726	6.168	3,447	4.124	4.017	4.550	5.149	4.957	6.743	3.346
1.168	1.763	0,622	2.786	4.782	3.397	3.994	1.921	1.714	2.373
3.238	4.557	**3,219**	3.335	0.641	10.599	10.406	11.662	0.222	10.237
0.283	0.163	0,814	2.302	1.101	2.715	0.470	0.462	0.814	2.980
3.054	1.277	3,099	1.934	0.206	1.929	0.575	1.089	0.289	1.435
3.483	3.547	3,129	5.710	3.334	3.645	5.478	**3.298**	4.686	3.469
0.668	1.180	3,188	2.429	4.009	3.122	2.252	1.105	4.255	2.085
2.576	2.268	3,545	1.127	2.069	3.408	3.349	3.863	2.491	3.414
4.385	1.285	**3,219**	1.937	1.812	2.162	2.081	4.421	4.249	4.599
1.589	1.276	2,751	0.819	2.093	2.700	2.421	2.740	2.224	2.820
0.706	1.523	4,851	1.602	4.022	1.399	1.671	2.287	4.115	1.108
1.721	3.194	1,051	3.420	1.406	3.575	1.586	**3.298**	1.712	3.696
1.535	1.701	1,466	1.192	2.600	3.875	2.265	1.995	1.767	3.724
3.876	1.856	1,723	1.872	2.278	1.143	1.079	2.902	1.891	2.860
0.737	2.047	**3,219**	2.488	1.351	1.041	2.934	2.882	1.617	1.052
2.750	5.298	2,372	5.287	5.913	4.634	4.520	3.012	4.673	3.123
1.373	1.996	3,664	1.678	3.197	1.797	2.731	**3.298**	2.728	1.343
3.386	1.849	**3,219**	5.218	6.036	2.054	6.604	2.182	1.310	2.984
4.755	3.972	1,879	3.576	2.127	2.750	1.792	1.623	2.187	3.749
2.650	2.213	1,241	2.986	2.135	1.215	1.608	1.562	1.126	1.524
5.571	3.181	**3,219**	5.382	3.060	3.755	1.035	4.237	5.737	5.339
1.530	2.504	2,470	2.068	1.122	0.344	3.872	1.045	3.311	1.349
4.779	4.420	3,471	4.447	0.445	4.719	3.270	**3.298**	4.179	3.091
2.452	4.650	**3,219**	2.951	4.005	0.832	2.911	2.574	2.371	1.803
2.565	2.414	0,923	2.062	5.526	2.385	1.990	2.036	2.973	2.421
1.439	3.829	1,334	1.294	1.279	2.422	2.949	2.741	1.932	2.659
3.888	1.524	3,675	4.748	7.131	7.411	7.808	1.854	5.252	5.882
1.603	1.507	4,001	2.180	1.244	1.084	2.942	1.930	2.045	1.612
2.633	1.371	1,907	2.073	1.416	1.304	2.665	3.206	1.354	1.596
2.086	1.962	1,252	1.197	1.661	1.713	2.182	**3.298**	2.399	2.791
2.800	1.987	**3,219**	1.832	6.257	1.129	6.075	7.053	1.242	6.120
7.423	6.601	6,400	3.976	3.149	1.643	7.398	7.141	4.436	6.879
3.786	6.453	**3,219**	6.418	6.050	5.496	3.591	6.079	1.401	3.806
1.755	6.641	1,837	5.535	3.645	5.206	3.588	**3.298**	3.233	1.775
0.804	2.132	5,803	3.424	2.305	3.475	7.773	7.824	2.168	4.732
1.661	1.418	2,400	3.917	4.567	1.186	1.240	3.133	1.511	1.656
4.292	4.003	3,284	4.179	3.924	4.342	4.589	**3.298**	1.153	4.109
4.955	2.839	4,372	3.730	3.567	3.045	3.825	5.077	3.874	2.255
4.301	1.327	**3,219**	2.704	3.954	2.647	4.671	2.970	1.283	2.873
2.509	1.469	3,747	3.180	7.432	4.313	7.123	4.382	7.261	4.588
1.275	9.904	1,865	1.178	6.441	3.053	1.436	**3.298**	0.179	1.579
4.694	3.156	**3,219**	7.665	6.024	4.361	4.524	2.119	6.514	2.655
0.705	3.267	**3,219**	0.557	2.272	2.904	1.237	2.449	1.013	2.028
2.262	4.162	3,531	1.048	1.417	1.594	3.558	1.702	1.956	1.286
3.973	3.493	1,691	3.246	2.600	4.683	3.667	4.641	3.274	4.739
1.411	1.568	0,709	1.908	2.580	1.461	2.729	**3.298**	2.888	0.146
2.416	1.431	**3,219**	1.198	1.046	2.869	6.104	3.508	4.971	6.288
3.240	1.273	**3,219**	1.785	3.923	4.030	2.579	4.832	3.118	4.303
1.458	2.949	2,079	3.588	1.777	3.941	1.778	1.587	1.203	1.796
4.904	5.356	5,279	5.169	10.262	5.529	10.492	**3.298**	5.913	10.542

Elaborado por: G. Cuenca

Viene...

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Imputación por Media

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.919	4.446	1,333	4.688	2.057	4.830	0.712	4.278	0.169	4.367
0.177	3.187	**3,219**	2.848	0.209	2.757	0.875	2.298	2.544	2.163
6.596	7.337	7,012	6.574	7.968	6.449	9.439	**3.298**	11.604	11.447
0.511	0.453	2,859	3.076	0.471	1.660	3.090	3.111	2.044	3.462
7.019	5.192	**3,219**	1.971	7.147	1.580	7.974	5.108	3.734	3.566
5.082	5.470	6,423	10.571	10.914	6.174	5.790	4.342	6.142	7.809
5.216	0.577	0,070	4.630	5.805	6.604	6.580	2.890	6.806	0.555
5.508	5.838	**3,219**	1.988	0.277	8.785	3.236	0.196	8.269	9.167
7.531	7.868	**3,219**	6.971	10.050	5.283	10.384	5.845	6.612	5.274
1.219	2.384	2,895	1.906	5.324	2.125	4.701	**3.298**	2.949	2.660
7.965	5.846	8,665	5.610	5.002	4.962	4.533	**3.298**	6.117	4.267
0.773	2.720	1,633	2.129	1.205	0.556	0.720	**3.298**	0.521	0.184
0.449	2.003	4,027	2.725	2.785	2.466	4.397	**3.298**	4.170	2.684
6.455	7.185	7,863	3.065	4.945	2.619	1.508	1.379	1.302	1.192
4.833	1.780	2,271	2.454	1.586	2.595	2.939	1.324	1.128	4.257
1.653	2.624	0,779	0.238	0.172	1.338	2.313	1.290	1.440	2.493
5.029	3.679	**3,219**	4.295	3.063	5.534	4.939	4.058	5.257	4.231
3.027	6.997	3,002	3.647	1.625	2.274	1.651	3.216	4.641	1.289
4.947	4.069	**3,219**	4.715	5.132	4.946	4.934	0.827	4.110	4.323
1.047	1.023	4,330	3.551	4.398	2.603	1.513	1.317	4.113	1.171
2.160	3.286	**3,219**	5.008	5.835	4.443	5.692	6.458	6.420	6.410
3.437	4.315	2,402	3.724	4.977	2.237	3.348	3.577	4.924	3.505
2.650	4.631	4,361	2.749	4.810	4.374	2.653	2.303	2.003	4.456
4.387	4.031	**3,219**	1.303	2.059	3.308	2.004	4.271	4.820	3.195
3.349	2.733	2,041	4.734	3.214	3.010	2.136	**3.298**	1.895	2.561
5.950	5.100	5,241	8.751	8.797	5.607	6.784	5.941	8.083	5.750
2.180	4.490	1,422	3.254	2.905	3.984	4.586	**3.298**	4.684	5.501
1.096	3.067	1,154	3.048	2.318	2.521	2.126	1.073	4.016	4.150
2.541	3.118	**3,219**	1.888	3.642	1.282	3.155	0.424	3.997	1.188
1.782	2.558	1,205	1.638	2.784	3.678	3.476	1.468	1.700	1.718
5.260	2.272	**3,219**	1.292	4.339	4.104	2.877	3.287	3.006	2.248
3.872	3.320	1,821	3.069	1.131	3.017	1.615	1.421	3.691	2.732
0.977	5.323	3,878	5.360	1.664	1.563	3.183	1.979	1.301	3.020
1.538	0.544	**3,219**	1.379	4.166	1.871	2.308	4.817	3.755	1.849
3.097	3.744	2,224	2.974	2.029	3.689	3.154	0.622	2.684	3.376
2.264	1.749	**3,219**	1.056	2.011	1.089	1.400	1.754	2.505	2.449
1.922	1.135	2,030	2.992	1.665	1.782	3.061	**3.298**	3.322	3.144
5.309	1.632	5,489	0.409	6.785	5.881	5.931	**3.298**	7.342	5.589
2.024	2.555	3,541	3.185	1.807	1.535	2.964	3.691	1.676	1.626
1.962	1.450	2,667	3.870	4.081	1.627	3.066	4.395	4.515	3.001
3.514	6.951	1,244	2.751	2.468	2.018	2.323	1.230	4.707	1.959
0.750	0.800	0,449	1.177	1.890	1.178	2.311	**3.298**	0.035	1.687
0.537	1.374	1,158	5.727	1.508	5.355	1.709	**3.298**	1.359	1.518
10.796	10.465	10,521	8.610	8.558	8.599	8.032	10.596	8.551	8.123
1.056	1.232	2,367	1.325	0.526	1.676	2.971	2.542	2.939	2.814
8.118	8.500	9,924	9.254	10.405	9.636	10.348	9.218	9.805	10.046
4.040	4.244	3,613	3.099	4.680	6.852	2.452	**3.298**	4.986	5.924
4.949	7.182	4,366	3.236	4.999	3.716	3.930	**3.298**	7.191	4.460
3.296	7.442	7,542	4.733	4.720	3.237	3.182	**3.298**	5.351	3.424
5.786	6.620	6,717	5.252	5.305	5.491	6.526	**3.298**	6.722	4.637

Tabla 4.34

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Imputación por Regresiòn

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
6.726	6.168	3,447	4.124	4.017	4.550	5.149	4,957	6.743	3.346
1.168	1.763	0,622	2.786	4.782	3.397	3.994	1,921	1.714	2.373
3.238	4.557	**2,849**	3.335	0.641	10.599	10.406	11,662	0.222	10.237
0.283	0.163	0,814	2.302	1.101	2.715	0.470	0,462	0.814	2.980
3.054	1.277	3,099	1.934	0.206	1.929	0.575	1,089	0.289	1.435
3.483	3.547	3,129	5.710	3.334	3.645	5.478	**2,386**	4.686	3.469
0.668	1.180	3,188	2.429	4.009	3.122	2.252	1,105	4.255	2.085
2.576	2.268	3,545	1.127	2.069	3.408	3.349	3,863	2.491	3.414
4.385	1.285	**1,403**	1.937	1.812	2.162	2.081	4,421	4.249	4.599
1.589	1.276	2,751	0.819	2.093	2.700	2.421	2,740	2.224	2.820
0.706	1.523	4,851	1.602	4.022	1.399	1.671	2,287	4.115	1.108
1.721	3.194	1,051	3.420	1.406	3.575	1.586	**1,102**	1.712	3.696
1.535	1.701	1,466	1.192	2.600	3.875	2.265	1,995	1.767	3.724
3.876	1.856	1,723	1.872	2.278	1.143	1.079	2,902	1.891	2.860
0.737	2.047	**1,057**	2.488	1.351	1.041	2.934	2,882	1.617	1.052
2.750	5.298	2,372	5.287	5.913	4.634	4.520	3,012	4.673	3.123
1.373	1.996	3,664	1.678	3.197	1.797	2.731	**1,374**	2.728	1.343
3.386	1.849	**6,399**	5.218	6.036	2.054	6.604	2,182	1.310	2.984
4.755	3.972	1,879	3.576	2.127	2.750	1.792	1,623	2.187	3.749
2.650	2.213	1,241	2.986	2.135	1.215	1.608	1,562	1.126	1.524
5.571	3.181	**3,909**	5.382	3.060	3.755	1.035	4,237	5.737	5.339
1.530	2.504	2,470	2.068	1.122	0.344	3.872	1,045	3.311	1.349
4.779	4.420	3,471	4.447	0.445	4.719	3.270	**1,207**	4.179	3.091
2.452	4.650	**1,098**	2.951	4.005	0.832	2.911	2,574	2.371	1.803
2.565	2.414	0,923	2.062	5.526	2.385	1.990	2,036	2.973	2.421
1.439	3.829	1,334	1.294	1.279	2.422	2.949	2,741	1.932	2.659
3.888	1.524	3,675	4.748	7.131	7.411	7.808	1,854	5.252	5.882
1.603	1.507	4,001	2.180	1.244	1.084	2.942	1,930	2.045	1.612
2.633	1.371	1,907	2.073	1.416	1.304	2.665	3,206	1.354	1.596
2.086	1.962	1,252	1.197	1.661	1.713	2.182	**2,601**	2.399	2.791
2.800	1.987	**6,554**	1.832	6.257	1.129	6.075	7,053	1.242	6.120
7.423	6.601	6,400	3.976	3.149	1.643	7.398	7,141	4.436	6.879
3.786	6.453	**2,226**	6.418	6.050	5.496	3.591	6,079	1.401	3.806
1.755	6.641	1,837	5.535	3.645	5.206	3.588	**6,003**	3.233	1.775
0.804	2.132	5,803	3.424	2.305	3.475	7.773	7,824	2.168	4.732
1.661	1.418	2,400	3.917	4.567	1.186	1.240	3,133	1.511	1.656
4.292	4.003	3,284	4.179	3.924	4.342	4.589	**4,007**	1.153	4.109
4.955	2.839	4,372	3.730	3.567	3.045	3.825	5,077	3.874	2.255
4.301	1.327	**1,673**	2.704	3.954	2.647	4.671	2,970	1.283	2.873
2.509	1.469	3,747	3.180	7.432	4.313	7.123	4,382	7.261	4.588
1.275	9.904	1,865	1.178	6.441	3.053	1.436	**19,618**	0.179	1.579
4.694	3.156	**1,429**	7.665	6.024	4.361	4.524	2,119	6.514	2.655
0.705	3.267	**3,688**	0.557	2.272	2.904	1.237	2,449	1.013	2.028
2.262	4.162	3,531	1.048	1.417	1.594	3.558	1,702	1.956	1.286
3.973	3.493	1,691	3.246	2.600	4.683	3.667	4,641	3.274	4.739
1.411	1.568	0,709	1.908	2.580	1.461	2.729	**5,832**	2.888	0.146
2.416	1.431	**3,952**	1.198	1.046	2.869	6.104	3,508	4.971	6.288
3.240	1.273	**3,411**	1.785	3.923	4.030	2.579	4,832	3.118	4.303
1.458	2.949	2,079	3.588	1.777	3.941	1.778	1,587	1.203	1.796
4.904	5.356	5,279	5.169	10.262	5.529	10.492	**5,243**	5.913	10.542

Continúa...

Viene...

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos de variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Imputación por Regresiòn

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.919	4.446	1,333	4.688	2.057	4.830	0.712	4,278	0.169	4.367
0.177	3.187	**2,689**	2.848	0.209	2.757	0.875	2,298	2.544	2.163
6.596	7.337	7,012	6.574	7.968	6.449	9.439	**9,492**	11.604	11.447
0.511	0.453	2,859	3.076	0.471	1.660	3.090	3,111	2.044	3.462
7.019	5.192	**6,713**	1.971	7.147	1.580	7.974	5,108	3.734	3.566
5.082	5.470	6,423	10.571	10.914	6.174	5.790	4,342	6.142	7.809
5.216	0.577	0,070	4.630	5.805	6.604	6.580	2,890	6.806	0.555
5.508	5.838	**0,853**	1.988	0.277	8.785	3.236	0,196	8.269	9.167
7.531	7.868	**6,397**	6.971	10.050	5.283	10.384	5,845	6.612	5.274
1.219	2.384	2,895	1.906	5.324	2.125	4.701	**3,003**	2.949	2.660
7.965	5.846	8,665	5.610	5.002	4.962	4.533	**2,572**	6.117	4.267
0.773	2.720	1,633	2.129	1.205	0.556	0.720	**0,919**	0.521	0.184
0.449	2.003	4,027	2.725	2.785	2.466	4.397	**4,489**	4.170	2.684
6.455	7.185	7,863	3.065	4.945	2.619	1.508	1,379	1.302	1.192
4.833	1.780	2,271	2.454	1.586	2.595	2.939	1,324	1.128	4.257
1.653	2.624	0,779	0.238	0.172	1.338	2.313	1,290	1.440	2.493
5.029	3.679	**3,562**	4.295	3.063	5.534	4.939	4,058	5.257	4.231
3.027	6.997	3,002	3.647	1.625	2.274	1.651	3,216	4.641	1.289
4.947	4.069	**4,993**	4.715	5.132	4.946	4.934	0,827	4.110	4.323
1.047	1.023	4,330	3.551	4.398	2.603	1.513	1,317	4.113	1.171
2.160	3.286	**4,052**	5.008	5.835	4.443	5.692	6,458	6.420	6.410
3.437	4.315	2,402	3.724	4.977	2.237	3.348	3,577	4.924	3.505
2.650	4.631	4,361	2.749	4.810	4.374	2.653	2,303	2.003	4.456
4.387	4.031	**2,075**	1.303	2.059	3.308	2.004	4,271	4.820	3.195
3.349	2.733	2,041	4.734	3.214	3.010	2.136	**2,078**	1.895	2.561
5.950	5.100	5,241	8.751	8.797	5.607	6.784	5,941	8.083	5.750
2.180	4.490	1,422	3.254	2.905	3.984	4.586	**3,642**	4.684	5.501
1.096	3.067	1,154	3.048	2.318	2.521	2.126	1,073	4.016	4.150
2.541	3.118	**1,129**	1.888	3.642	1.282	3.155	0,424	3.997	1.188
1.782	2.558	1,205	1.638	2.784	3.678	3.476	1,468	1.700	1.718
5.260	2.272	**3,347**	1.292	4.339	4.104	2.877	3,287	3.006	2.248
3.872	3.320	1,821	3.069	1.131	3.017	1.615	1,421	3.691	2.732
0.977	5.323	3,878	5.360	1.664	1.563	3.183	1,979	1.301	3.020
1.538	0.544	**1,922**	1.379	4.166	1.871	2.308	4,817	3.755	1.849
3.097	3.744	2,224	2.974	2.029	3.689	3.154	0,622	2.684	3.376
2.264	1.749	**1,348**	1.056	2.011	1.089	1.400	1,754	2.505	2.449
1.922	1.135	2,030	2.992	1.665	1.782	3.061	**4,910**	3.322	3.144
5.309	1.632	5,489	0.409	6.785	5.881	5.931	**6,289**	7.342	5.589
2.024	2.555	3,541	3.185	1.807	1.535	2.964	3,691	1.676	1.626
1.962	1.450	2,667	3.870	4.081	1.627	3.066	4,395	4.515	3.001
3.514	6.951	1,244	2.751	2.468	2.018	2.323	1,230	4.707	1.959
0.750	0.800	0,449	1.177	1.890	1.178	2.311	**0,697**	0.035	1.687
0.537	1.374	1,158	5.727	1.508	5.355	1.709	**5,407**	1.359	1.518
10.796	10.465	10,521	8.610	8.558	8.599	8.032	10,596	8.551	8.123
1.056	1.232	2,367	1.325	0.526	1.676	2.971	2,542	2.939	2.814
8.118	8.500	9,924	9.254	10.405	9.636	10.348	9,218	9.805	10.046
4.040	4.244	3,613	3.099	4.680	6.852	2.452	**3,192**	4.986	5.924
4.949	7.182	4,366	3.236	4.999	3.716	3.930	**4,846**	7.191	4.460
3.296	7.442	7,542	4.733	4.720	3.237	3.182	**4,782**	5.351	3.424
5.786	6.620	6,717	5.252	5.305	5.491	6.526	**5,493**	6.722	4.637

En la Tabla 4.35 se realiza una comparación entre el valor real y el valor con imputación por la media y regresión.

Tabla 4.35

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Comparación de los Métodos de Imputación

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

25% de datos completados en X_3 por la Media

Dato Observado	Resultado de Imputación por Media	Error |Dato Observado – Resultado de Imputación por Media|
2.851	3.219	0,368
1.414	3.219	1,805
1.069	3.219	2,150
6.462	3.219	3,243
3.914	3.219	0,695
1.131	3.219	2,088
6.562	3.219	3,343
2.254	3.219	0,965
1.689	3.219	1,530
1.432	3.219	1,787
3.693	3.219	0,474
3.960	3.219	0,741
3.420	3.219	0,201
2.683	3.219	0,536
6.730	3.219	3,511
0.860	3.219	2,359
6.406	3.219	3,187
3.578	3.219	0,359
5.157	3.219	1,938
4.083	3.219	0,864
2.061	3.219	1,158
1.148	3.219	2,071
3.359	3.219	0,140
1.913	3.219	1,306
1.351	3.219	1,868

25% de datos completados en X_3 por Regresión

Dato Observado	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
2.851	2.849	0,002
1.414	1.403	0,011
1.069	1.057	0,012
6.462	6.399	0,063
3.914	3.909	0,005
1.131	1.098	0,033
6.562	6.554	0,008
2.254	2.226	0,028
1.689	1.673	0,016
1.432	1.429	0,003
3.693	3.688	0,005
3.960	3.952	0,008
3.420	3.411	0,009
2.683	2.689	0,006
6.730	6.713	0,017
0.860	0.853	0,007
6.406	6.397	0,009
3.578	3.562	0,016
5.157	4.993	0,164
4.083	4.052	0,031
2.061	2.075	0,014
1.148	1.129	0,019
3.359	3.347	0,012
1.913	1.922	0,009
1.351	1.348	0,003

Elaborado por: G. Cuenca

Continúa…

Viene...

Efectos de la Imputación en el análisis de datos multivariados
Variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Comparación de los Métodos de Imputación
Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

25% de datos completados en X_8 por la Media

Dato Observado	Resultado de Imputación por Media	Error |Dato Observado – Resultado de Imputación por Media|
2.390	3.298	0,908
1.060	3.298	2,238
1.383	3.298	1,915
1.219	3.298	2,079
2.582	3.298	0,716
5.997	3.298	2,699
3.952	3.298	0,654
19.664	3.298	16,366
5.859	3.298	2,561
5.255	3.298	1,957
9.518	3.298	6,22
2.947	3.298	0,351
2.566	3.298	0,732
0.929	3.298	2,369
4.580	3.298	1,282
2.080	3.298	1,218
3.767	3.298	0,469
4.930	3.298	1,632
6.314	3.298	3,016
0.704	3.298	2,594
5.413	3.298	2,115
3.183	3.298	0,115
4.859	3.298	1,561
4.800	3.298	1,502
5.525	3.298	2,227

25% de datos completados en X_8 por Regresión

Dato Observado	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
2.390	2.386	0,004
1.060	1.102	0,042
1.383	1.374	0,009
1.219	1.207	0,012
2.582	2.601	0,019
5.997	6.003	0,006
3.952	4.007	0,055
19.664	19.618	0,046
5.859	5.832	0,027
5.255	5.243	0,012
9.518	9.492	0,026
2.947	3.003	0,056
2.566	2.572	0,006
0.929	0.919	0,010
4.580	4.489	0,091
2.080	2.078	0,002
3.767	3.642	0,125
4.930	4.910	0,020
6.314	6.289	0,025
0.704	0.697	0,007
5.413	5.407	0,006
3.183	3.192	0,009
4.859	4.846	0,013
4.800	4.782	0,018
5.525	5.493	0,032

Elaborado por: G. Cuenca

Se puede notar, por medio de la Tabla 4.35 que la diferencia en valor absoluto entre el dato observado y el estimado de cada variable es menor en el Método de Imputación por Regresión.

CUADRO 4.44

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Imputación por la Media y Regresión

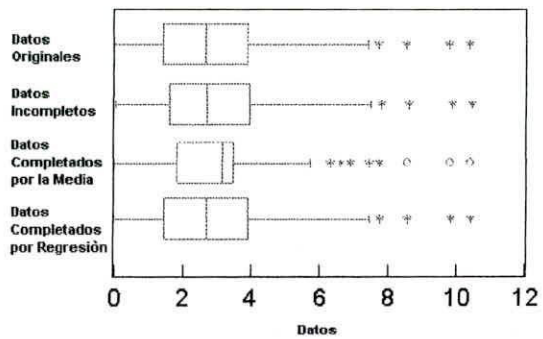
Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Variable* X_3" y "*Variable* X_8"

Estimadores "*Variable* X_3"

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		100	75	100	100
Media		3,206	3,219	3,219	3,201
Mediana		2,801	2,751	3,219	2,800
Moda		0,070	0,070	3,220	0,070
Varianza		4,528	4,901	3,663	4,518
Desviación Estándar		2,128	2,214	1,914	2,126
Error Estándar		0,213	0,256	0,191	0,213
Coeficiente de Asimetría		1,194	1,294	1,486	1,198
Curtosis		1,351	1,630	3,139	1,372
Rango		10,450	10,450	10,450	10,450
Mínimo		0,070	0,070	0,070	0,070
Máximo		10,520	10,520	10,520	10,520
Percentiles	25	1,508	1,633	1,886	1,508
	50	2,801	2,751	3,219	2,800
	75	4,021	4,027	3,596	4,021

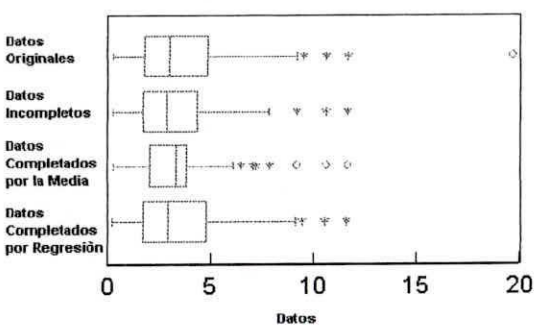
Diagrama de Cajas "*Variable* X_3"



Estimadores "*Variable* X_8"

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		100	75	100	100
Media		3,588	3,298	3,298	3,585
Mediana		2,959	2,882	3,298	2,987
Moda		0,200	0,200	3,300	0,200
Varianza		7,626	5,164	3,860	7,599
Desviación Estándar		2,762	2,273	1,965	2,757
Error Estándar		0,276	0,262	0,197	0,276
Coeficiente de Asimetría		2,576	1,484	1,704	2,573
Curtosis		11,269	2,742	4,597	11,234
Rango		19,470	11,470	11,470	19,420
Mínimo		0,200	0,200	0,200	0,200
Máximo		19,660	11,660	11,660	19,620
Percentiles	25	1,715	1,623	1,983	1,715
	50	2,959	2,882	3,298	2,987
	75	4,813	4,382	3,820	4,808

Diagrama de Cajas "*Variable* X_8"



Elaborado por: G. Cuenca

Al realizar la imputación por la media y regresión se obtuvieron los siguientes resultados en la variable X_3 (Ver Cuadro 4.44):

El valor de la media de los "datos completados" por *la media* aumenta, comparándolo con los "datos originales" y completados por *regresión.*

El valor de la varianza de los "datos completados" por la *media* disminuye de 4.528 a 3.663, mientras que en los datos completados por regresión este valor se incrementa a 4.518, comparándolo con el valor anterior y es muy cercano al valor de la varianza de los datos originales.

Mientras que en la variable X_8, el valor de la media de los "datos completados" por *la media* aumenta, comparándolo con los "datos originales" y completados por *regresión.*

El valor de la varianza de los "datos completados" por la *media* disminuye de 7.626 a 3.860. Esta variable presenta valores atípicos.

El vector de medias con veinticinco datos completados por la media en X_3 y veinticinco en X_8 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \\ \overline{X}_7 \\ \overline{X}_8 \\ \overline{X}_9 \\ \overline{X}_{10} \end{pmatrix} = \begin{pmatrix} 3.164 \\ 3.445 \\ 3.219 \\ 3.350 \\ 3.614 \\ 3.391 \\ 3.741 \\ 3.298 \\ 3.526 \\ 3.532 \end{pmatrix}$$

Mientras que el vector de medias con veinticinco datos completados por la regresión en X_3 y veinticinco en X_8 es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \\ \overline{X}_7 \\ \overline{X}_8 \\ \overline{X}_9 \\ \overline{X}_{10} \end{pmatrix} = \begin{pmatrix} 3.164 \\ 3.445 \\ 3.201 \\ 3.350 \\ 3.614 \\ 3.391 \\ 3.741 \\ 3.585 \\ 3.526 \\ 3.532 \end{pmatrix}$$

El efecto que causa en la *matriz de varianzas y covarianzas* y *matriz de correlaciones*, el completar 10% de datos faltantes en una matriz de tamaño 100, por medio de la imputación por media y regresión, se presenta en el Cuadro 4.45.

CUADRO 4.45

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Imputación por Media y Regresión

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas
(Datos Originales)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	4.386									
X_2	2.700	4.854								
X_3	2.701	2.336	4.528							
X_4	2.165	2.247	2.041	3.978						
X_5	2.780	2.346	3.072	2.968	6.029					
X_6	2.252	1.997	1.489	2.269	2.240	4.084				
X_7	2.706	1.857	2.925	2.329	3.696	2.695	5.563			
X_8	1.637	2.954	2.226	1.497	2.897	2.173	3.059	7.626		
X_9	3.019	2.133	2.366	2.365	3.272	2.508	3.039	1.543	5.322	
X_{10}	2.552	2.045	2.371	2.044	2.521	3.192	3.685	2.716	2.939	5.072

Matriz de Varianzas y Covarianzas
25% Datos Completados por Media en "*Variable* X_3" y 25% en "*Variable* X_8"

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	4.386									
X_2	2.700	4.854								
X_3	2.402	2.182	3.663							
X_4	2.165	2.247	1.826	3.978						
X_5	2.780	2.346	2.401	2.968	6.029					
X_6	2.252	1.997	1.509	2.269	2.240	4.084				
X_7	2.706	1.857	2.224	2.329	3.696	2.695	5.563			
X_8	1.611	1.419	1.629	1.522	1.735	1.728	2.697	3.860		
X_9	3.019	2.133	2.404	2.365	3.272	2.508	3.039	1.139	5.322	
X_{10}	2.552	2.045	2.138	2.044	2.521	3.192	**3.685**	2.296	2.939	5.072

Matriz de Varianzas y Covarianzas
25% Datos Completados por Regresión en "*Variable* X_3" y 25% en "*Variable* X_8"

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	4.386									
X_2	2.700	4.854								
X_3	2.697	2.335	4.518							
X_4	2.165	2.247	2.037	3.978						
X_5	2.780	2.346	3.065	2.968	6.029					
X_6	2.252	1.997	1.487	2.269	2.240	4.084				
X_7	2.706	1.857	2.919	2.329	3.696	2.695	5.563			
X_8	1.639	2.950	2.228	1.498	2.895	2.172	3.056	7.599		
X_9	3.019	2.133	2.367	2.365	3.272	2.508	3.039	1.536	5.322	
X_{10}	2.552	2.045	2.369	2.044	2.521	3.192	3.685	2.713	2.939	5.072

Elaborado por: G. Cuenca

CUADRO 4.46

Efectos de la Imputación en el análisis de datos multivariados

Variables aleatorias dependientes con distribución Exponencial $\beta = 4$

Método de Imputación por Media y Regresión

Tamaño de muestra n=100 y 5% de datos faltantes en la matriz

Matriz de Correlaciones
(Datos Originales)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000									
X_2	0.585	1.000								
X_3	0.606	0.498	1.000							
X_4	0.518	0.511	0.481	1.000						
X_5	0.541	0.434	0.588	0.606	1.000					
X_6	0.532	0.448	0.346	0.563	0.451	1.000				
X_7	0.548	0.357	0.583	0.495	0.638	0.565	1.000			
X_8	0.283	0.486	0.379	0.272	0.427	0.389	0.470	1.000		
X_9	0.625	0.420	0.482	0.514	0.578	0.538	0.559	0.242	1.000	
X_{10}	0.541	0.412	0.495	0.455	0.456	0.701	0.694	0.437	0.566	1.000

Matriz de Correlaciones
25% Datos Completados por Media en "*Variable* X_3" y 25% en "*Variable* X_8"

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000									
X_2	0.585	1.000								
X_3	0.599	0.517	1.000							
X_4	0.518	0.511	0.478	1.000						
X_5	0.541	0.434	0.511	0.606	1.000					
X_6	0.532	0.448	0.390	0.563	0.451	1.000				
X_7	0.548	0.357	0.493	0.495	0.638	0.565	1.000			
X_8	0.392	0.328	0.433	0.388	0.360	0.435	0.582	1.000		
X_9	0.625	0.420	0.544	0.514	0.578	0.538	0.559	0.251	1.000	
X_{10}	0.541	0.412	0.496	0.455	0.456	0.701	0.694	0.519	0.566	1.000

Matriz de Correlaciones
25% Datos Completados por Regresión en "*Variable* X_3" y 25% en "*Variable* X_8"

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000									
X_2	0.585	1.000								
X_3	0.606	0.499	1.000							
X_4	0.518	0.511	0.480	1.000						
X_5	0.541	0.434	0.587	0.606	1.000					
X_6	0.532	0.448	0.346	0.563	0.451	1.000				
X_7	0.548	0.357	0.582	0.495	0.638	0.565	1.000			
X_8	0.284	0.486	0.380	0.272	0.428	0.390	0.470	1.000		
X_9	0.625	0.420	0.483	0.514	0.578	0.538	0.559	0.242	1.000	
X_{10}	0.541	0.412	0.495	0.455	0.456	0.701	0.694	0.437	0.566	1.000

Elaborado por: G. Cuenca

Se puede apreciar en el Cuadro 4.45, que los únicos valores que cambian son las covarianzas de la variable X_3 y X_8 con las demás variables, donde la covarianza entre X_3 y X_5 , disminuye de 3.072 a 2.401.

En la matriz de varianzas y covarianzas de los datos completados por regresión, el valor de las covarianzas de variable X_3 y X_8 con las demás variables se incrementa, comparándolo con la matriz de varianzas y covarianzas de los "datos completados" por *la media.*

Por otro lado, analizando el efecto que causa en la matriz de correlaciones, podemos apreciar en le Cuadro 4.46 que también los únicos valores que cambian son los de la correlación de X_3 y X_8 con las demás variables, puesto que a estas variables se les completó datos por medio de los métodos de imputación; donde la mayor correlación se da entre las variables X_6 y X_{10}, es decir 0.701, seguida por 0.694 entre las variables X_7 y X_{10}. En la matriz de correlaciones con 25% de datos completados por la media en X_3 y 25% en X_8, la correlación entre X_3 y X_5 disminuye de 0.588 a 0.511, mientras que en la matriz de datos completados por regresión, este valor es 0.587, es decir tiende al valor observado.

CONCLUSIONES

Las conclusiones presentadas a continuación se derivan de los análisis realizados en los capítulos anteriores de esta investigación, basados en los Efectos de la Imputación en el análisis de datos multivariados. Para realizar este análisis se realizaron simulaciones en diferentes tamaños de muestra: 30, 50 y 100.

Después de indicar la base de este estudio, se presentan las conclusiones de cuando se trabaja con matrices de datos con variables aleatorias independientes y dependientes.

Cuando se trabaja con matrices de datos con variables aleatorias independientes se obtienen las siguientes conclusiones:

1. 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.

2. No existe gran diferencia en la *matriz de varianzas y covarianzas* y *de correlaciones*, cuando se completa datos por imputación por media y regresión, si el tamaño de muestra es n=30 con el 2% de datos faltantes.

3. 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.

4. Cuando se trabaja en matrices de datos con variables aleatorias independientes, el *Método de Imputación por Regresión* brinda resultados de predicción que no tienden al "dato observado", pero están más cercanos a los valores que estima el *Método de Imputación por Media*.

Cuando se trabaja con matrices de datos con variables aleatorias dependientes se obtienen las siguientes conclusiones:

5. A diferencia de cuando se trabaja con muestras tomadas de poblaciones independientes, 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.

6. Si se tiene una matriz de datos cuyas columnas son muestras tomadas de poblaciones poisson, dependientes e idénticamente distribuìdas, con tamaño de muestra n=100 con el 10% de datos faltantes, la *matriz de varianzas y covarianzas* tampoco se ve mayormente afectada.

7. Si la cantidad de datos faltantes es del 10%, para un tamaño de muestra n=100, la matriz de varianzas y covarianzas de "datos completados" por la media se ve afectada, ya que las covarianzas de la variables, a las que se les completó datos, varían.

8. La *matriz de varianzas y covarianzas* de datos completados por regresión no se ve afectada, ya que las covarianzas de las variables, a las que se les completó datos, tienden a los datos observados.

9. El *Método de Imputación por Media*, disminuye el valor de la varianza muestral de la variable, puesto que en el lugar del dato faltante se coloca el promedio de la variable con datos incompletos

10. 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.
11. La diferencia, en valor absoluto, entre el dato observado y el resutado de predicción, es menor cuando se imputa utilizando el método de regresión, más aún si se trabaja con matrices de datos con variables aleatorias dependientes.

RECOMENDACIONES

1. Antes de usar algún método de imputación, se debe obtener la *matriz de varianzas y covarianzas* y *matriz de correlaciones*, para de esta manera, conocer si las variables investigadas son o no independientes, utilizando por ejemplo el Método de Barlett.
2. Se recomienda utilizar el *Método de Eliminación por Filas*, cuando la cantidad de datos faltantes en un matriz es menor o igual al 2%, además es preferible que los datos faltantes estén en la misma fila.
3. Si la cantidad de datos faltantes, en una matriz de datos con muestras tomadas de poblaciones independientes es mayor al 2%, es recomendable utilizar algún método de imputación para estimar estos valores faltantes.
4. Si la matriz de datos contiene variables aleatorias independientes, se puede utilizar cualquiera de los dos métodos de imputación estudiados, pero debe recordarse que el método de *imputación por regresión* brinda resultados de predicción para cada uno de los datos faltantes, en cambio el método de imputación por la media nos da un solo valor.
5. Si la matriz de datos contiene variables aleatorias dependientes, es preferible utilizar el *Método de Imputación por Regresión,* puesto que, por medio de este método, los resultados de predicción tienden al valor observado.

6. Si los datos faltantes se encuentran solo en un ente investigado, con esto no es posible encontrar suficientes datos para calcular la ecuación de predicción inicial en el método de imputación por regresión. En esta situación se puede comenzar por usar la imputación por la media y luego usar *imputación por regresión* para las siguientes iteraciones.
7. Es preferible utilizar algún método de imputación antes que el *Método de eliminación por filas*, ya que si se eliminan filas de algún ente investigado se pierden datos de las otras características investigadas.
8. No se debe abusar de los *métodos de imputación*, debido a que realmente no se aumenta la información disponible sino que se genera a partir de la información que se posee.
9. Una idea básica que se debe de tener es que la imputación no sustituye ni descuida alguna fase previa, tal como la recolección de datos y digitalización. Hay que intentar obtener el dato original de las distintas variables por todos los medios disponibles y en el caso de no obtenerlo se recurrirá a la imputación de datos.

ANEXOS

Anexo 1

Caso Real

El Cuestionario y la Población Objetivo

A fin de utilizar los métodos de imputación estudiados anteriormente se procede a realizar el análisis con una base de datos proporcionada por el Centro de Investigaciones Estadísticas de la ESPOL, en esta base de datos la población objetivo fueron los Ejecutivos de las empresas que se encuentran en la ciudad de Guayaquil y que tienen ciertos cargos especiales como Presidente, Gerente general, Gerente de división, Director de Recursos Humanos, Accionista u otros que se encuentren a cargo del personal, a los que se les realizó un cuestionario para evaluar al profesional politécnico.

El Cuestionario proporcionado se divide en tres secciones, la primera subsección es *Acerca del entrevistado*, con este se pretende obtener información acerca de su nivel de instrucción, cargo que ocupa dentro de la organización, su título más alto y dónde lo obtuvo.

La segunda subsección es *Acerca de la organización*, en esta se puede obtener especificaciones tales como qué actividades tiene la organización, qué tipo de compañía es, donde está localizada, etc., esta tiene diferentes opciones de respuestas que están codificadas de 1 a 10 en unas como máximo y de 1 a 2 opciones como mínimo. Presentan también opciones de

respuestas que constan de: "Si", "No" y "No conozco tal Opción" en uno de los casos.

La tercera subsección es acerca *De los Profesionales Politécnicos* y a su vez esta se divide en dos partes. El primero especifica la "*Formación general de los Profesionales Politécnicos*" y la segunda con respecto a la "*Comunicación y Formación Específicas de los Profesionales Politécnicos*", en esta subsección las opciones de respuesta presentan *Escala Likert* que va desde cero a cinco, donde cero significa estar en completo desacuerdo con la misma y cinco completo acuerdo.

En general el cuestionario tuvo de 46 características que evaluó al Profesional Politécnico.

Del total de cuarenta y seis variables que se estudiaron se han seleccionado veinte y seis tomando en cuenta que son variables cuantitativas o cualitativas ordinales, puesto que los métodos de imputación estudiados trabajan con variables aleatorias cuantitativas.

Entonces la matriz de datos que se utiliza tiene 209 filas (número de entrevistados) y 26 columnas (número de variables). Las variables que integran la matriz son las siguientes:

X_1: *"Son personas con capacidad de análisis para llegar a conclusiones válidas, bajo distintas circunstancias"*

X_2: *"Tienen desarrollado su Pensamiento Crítico"*

X_3: *"Tienen marcado estilo de ver el mundo"*

X_4: *"Tienen capacidad para manejar los retos e innovaciones"*

X_5: *"Tienen competencia para formularse sus propias preguntas"*

X_6: *"Tienen habilidad para aprender por cuenta propia y por tanto mantenerse actualizados en los desarrollos que con el paso del tiempo, se dan en su área de competencia "*

X_7: *"Tienen habilidad para tomar decisiones oportunas"*

X_8: *"Saben trabajar en Equipo".*

X_9: *"Saben desarrollar actividades conjuntas con profesionales de áreas diferentes a la suya".*

X_{10}: *"Tienen claros propósitos de superación, esto es: tenacidad y estrategia".*

X_{11}: *"Tienen altos valores éticos y morales".*

X_{12}: *"Son gestores tecnológicos, esto es, son capaces de usar la tecnología que está a la mano"*

X_{13}: *"Se muestran siempre interesados y curiosos"*

X_{14}: "Su *formación es comparable a la de profesionales Extranjeros"*

X_{15}: *"Su presentación y comportamiento personal son siempre adecuados para la ocasión"*

X_{16}: "*Su proceso de ascenso en el organigrama de la Organización es notable*"

X_{17}: "*Son buenos comunicadores en forma oral*"

X_{18}: "*Son buenos comunicadores en forma escrita*"

X_{19}: "*Son personas que fácilmente se relacionan con terceros*"

X_{20}: "*Combinan de la mejor manera lo teórico con lo práctico*"

X_{21}: "*Son altamente capacitados para llevar a cabo Análisis Cuantitativos*"

X_{22}: "*Tienen alta compresión de los principios Físicos y Naturales*"

X_{23}: "*Sólida formación en Informática*"

X_{24}: "*Manejan los principios fundamentales de Administración*"

X_{25}: "*Muestran clara sensibilidad Social y Humana*"

X_{26}: "*Poseen el nivel de Inglés adecuado para utilizarlo de la manera requerida por sus actividades en la Organización*"

Implementación del Método de Eliminación por Filas y de los Métodos de Imputación

Se supone que la matriz de datos tiene 5% de valores faltantes, los cuales recayeron en las variables, "Su *formación es comparable a la de profesionales extranjeros*" y "*Su proceso de ascenso en el organigrama de la organización es notable*". Nótese que el 5% de datos faltantes constituyen 271 datos faltantes en la matriz, es decir

para este caso; 136 en la variable "Su *formación es comparable a la de profesionales extranjeros"* y 136 en la variable *"Su proceso de ascenso en el organigrama de la organización es notable"*. Donde los datos faltantes pueden recaer en la misma fila.

El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \\ \overline{X}_7 \\ \overline{X}_8 \\ \overline{X}_9 \\ \overline{X}_{10} \\ \overline{X}_{11} \\ \overline{X}_{12} \\ \overline{X}_{13} \\ \overline{X}_{14} \\ \overline{X}_{15} \\ \overline{X}_{16} \\ \overline{X}_{17} \\ \overline{X}_{18} \\ \overline{X}_{19} \\ \overline{X}_{20} \\ \overline{X}_{21} \\ \overline{X}_{22} \\ \overline{X}_{23} \\ \overline{X}_{24} \end{pmatrix} = \begin{pmatrix} 4.091 \\ 3.962 \\ 3.746 \\ 4.005 \\ 3.943 \\ 4.278 \\ 3.789 \\ 3.852 \\ 3.861 \\ 4.124 \\ 4.378 \\ 4.311 \\ 4.139 \\ 3.938 \\ 3.938 \\ 3.770 \\ 3.431 \\ 3.699 \\ 3.560 \\ 3.952 \\ 4.225 \\ 4.134 \\ 4.254 \\ 3.584 \\ 3.856 \\ 3.196 \end{pmatrix}$$

Método de Eliminación por Filas

Puesto que los datos faltantes recayeron en las variables "Su *formación es comparable a la de profesionales extranjeros"* y "*Su proceso de ascenso en el organigrama de la organización es notable"*, se procede a prescindir de 108 filas, es decir el total de filas eliminadas.

El vector de medias para las ciento un filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \\ \overline{X}_7 \\ \overline{X}_8 \\ \overline{X}_9 \\ \overline{X}_{10} \\ \overline{X}_{11} \\ \overline{X}_{12} \\ \overline{X}_{13} \\ \overline{X}_{14} \\ \overline{X}_{15} \\ \overline{X}_{16} \\ \overline{X}_{17} \\ \overline{X}_{18} \\ \overline{X}_{19} \\ \overline{X}_{20} \\ \overline{X}_{21} \\ \overline{X}_{22} \\ \overline{X}_{23} \\ X_{24} \end{pmatrix} = \begin{pmatrix} 2.010 \\ 1.923 \\ 1.871 \\ 1.938 \\ 1.923 \\ 2.105 \\ 1.866 \\ 1.847 \\ 1.837 \\ 2.014 \\ 2.134 \\ 2.110 \\ 2.005 \\ 1.928 \\ 1.933 \\ 1.828 \\ 1.651 \\ 1.780 \\ 1.722 \\ 1.919 \\ 2.081 \\ 2.000 \\ 2.086 \\ 1.785 \\ 1.852 \\ 1.574 \end{pmatrix}$$

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

Ahora analicemos el efecto que causa en la *matriz de varianzas y covarianzas*, y *matriz de correlaciones*, la eliminación de ciento ocho filas, tamaño de muestra n=209.

En las ocho hojas siguientes, se muestra la matriz de varianzas y covarianzas y de correlaciones de los datos originales y de los datos con filas eliminadas.

Matriz de Varianzas y Covarianzas (Datos Originales)

Tamaño de muestra n=209

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	0.679														
X_2	0.441	0.739													
X_3	0.312	0.322	1.007												
X_4	0.432	0.414	0.328	0.880											
X_5	0.419	0.402	0.283	0.481	0.679										
X_6	0.374	0.352	0.253	0.427	0.338	0.721									
X_7	0.399	0.372	0.268	0.515	0.435	0.333	0.773								
X_8	0.403	0.403	0.318	0.414	0.352	0.282	0.416	1.127							
X_9	0.311	0.355	0.229	0.380	0.333	0.250	0.350	0.749	0.947						
X_{10}	0.364	0.404	0.238	0.523	0.430	0.384	0.406	0.379	0.388	0.706					
X_{11}	0.235	0.231	0.231	0.354	0.286	0.298	0.282	0.388	0.332	0.309	0.630				
X_{12}	0.303	0.315	0.204	0.450	0.345	0.375	0.345	0.282	0.289	0.365	0.339	0.600			
X_{13}	0.329	0.337	0.338	0.374	0.301	0.413	0.366	0.309	0.269	0.353	0.269	0.385	0.649		
X_{14}	0.438	0.416	0.374	0.423	0.424	0.416	0.376	0.351	0.309	0.354	0.370	0.394	0.374	0.943	
X_{15}	0.376	0.401	0.388	0.390	0.347	0.416	0.357	0.414	0.381	0.460	0.341	0.332	0.388	0.448	1.001
X_{16}	0.362	0.352	0.206	0.453	0.362	0.324	0.389	0.384	0.357	0.356	0.255	0.331	0.345	0.414	0.389
X_{17}	0.278	0.305	0.244	0.363	0.323	0.279	0.370	0.338	0.373	0.321	0.245	0.197	0.252	0.320	0.431
X_{18}	0.249	0.296	0.390	0.357	0.295	0.204	0.326	0.301	0.285	0.297	0.268	0.219	0.278	0.347	0.375
X_{19}	0.266	0.300	0.215	0.387	0.330	0.214	0.344	0.482	0.516	0.339	0.259	0.234	0.234	0.275	0.386
X_{20}	0.312	0.296	0.243	0.462	0.358	0.311	0.442	0.339	0.320	0.381	0.259	0.303	0.305	0.348	0.372
X_{21}	0.297	0.364	0.331	0.364	0.335	0.336	0.288	0.255	0.243	0.357	0.299	0.324	0.353	0.355	0.375
X_{22}	0.257	0.274	0.260	0.302	0.291	0.270	0.245	0.241	0.235	0.277	0.238	0.256	0.255	0.340	0.326
X_{23}	0.241	0.236	0.267	0.335	0.284	0.285	0.270	0.144	0.170	0.324	0.288	0.339	0.306	0.309	0.309
X_{24}	0.298	0.369	0.182	0.440	0.413	0.260	0.474	0.385	0.375	0.417	0.302	0.308	0.260	0.320	0.387
X_{25}	0.229	0.321	0.228	0.304	0.271	0.232	0.229	0.397	0.326	0.340	0.247	0.175	0.246	0.279	0.366
X_{26}	0.203	0.229	0.189	0.350	0.237	0.272	0.258	0.332	0.263	0.245	0.147	0.203	0.199	0.224	0.267

Continúa…

Viene...

Matriz de Varianzas y Covarianzas (Datos Originales)
Tamaño de muestra n=209

Variables	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
X_{16}	0.812										
X_{17}	0.325	0.823									
X_{18}	0.229	0.467	0.769								
X_{19}	0.336	0.445	0.371	0.825							
X_{20}	0.311	0.352	0.332	0.392	0.642						
X_{21}	0.292	0.220	0.308	0.244	0.352	0.589					
X_{22}	0.281	0.216	0.276	0.199	0.271	0.383	0.559				
X_{23}	0.284	0.145	0.154	0.098	0.243	0.347	0.307	0.729			
X_{24}	0.351	0.377	0.268	0.364	0.432	0.306	0.234	0.332	0.946		
X_{25}	0.255	0.312	0.250	0.379	0.306	0.268	0.250	0.142	0.368	0.652	
X_{26}	0.358	0.290	0.228	0.265	0.255	0.215	0.171	0.195	0.332	0.177	0.995

Matriz de Correlaciones (Datos Originales)

Tamaño de muestra n=209

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	1.000														
X_2	0.623	1.000													
X_3	0.377	0.373	1.000												
X_4	0.559	0.513	0.349	1.000											
X_5	0.616	0.567	0.343	0.622	1.000										
X_6	0.534	0.482	0.297	0.536	0.483	1.000									
X_7	0.551	0.492	0.304	0.625	0.600	0.446	1.000								
X_8	0.461	0.442	0.298	0.416	0.402	0.313	0.445	1.000							
X_9	0.387	0.425	0.235	0.417	0.416	0.303	0.410	0.725	1.000						
X_{10}	0.525	0.559	0.283	0.664	0.621	0.538	0.550	0.425	0.474	1.000					
X_{11}	0.359	0.338	0.290	0.475	0.437	0.443	0.404	0.460	0.429	0.463	1.000				
X_{12}	0.475	0.473	0.263	0.620	0.540	0.570	0.506	0.343	0.383	0.561	0.551	1.000			
X_{13}	0.495	0.487	0.418	0.495	0.454	0.604	0.517	0.362	0.344	0.521	0.421	0.616	1.000		
X_{14}	0.548	0.498	0.383	0.465	0.530	0.505	0.441	0.341	0.327	0.434	0.480	0.524	0.478	1.000	
X_{15}	0.456	0.467	0.386	0.415	0.421	0.490	0.406	0.390	0.391	0.547	0.429	0.428	0.482	0.461	1.000
X_{16}	0.488	0.454	0.228	0.536	0.487	0.423	0.491	0.401	0.407	0.470	0.357	0.475	0.475	0.472	0.432
X_{17}	0.372	0.391	0.268	0.427	0.432	0.362	0.464	0.351	0.422	0.421	0.340	0.281	0.345	0.363	0.475
X_{18}	0.344	0.393	0.442	0.434	0.408	0.274	0.422	0.324	0.334	0.404	0.385	0.323	0.393	0.407	0.428
X_{19}	0.356	0.385	0.236	0.454	0.441	0.278	0.431	0.501	0.583	0.444	0.359	0.332	0.321	0.312	0.425
X_{20}	0.473	0.430	0.302	0.614	0.542	0.458	0.627	0.399	0.411	0.566	0.406	0.489	0.472	0.447	0.464
X_{21}	0.469	0.553	0.430	0.506	0.530	0.516	0.427	0.313	0.325	0.553	0.491	0.545	0.572	0.477	0.488
X_{22}	0.417	0.427	0.347	0.431	0.473	0.426	0.372	0.304	0.323	0.440	0.400	0.442	0.424	0.468	0.435
X_{23}	0.343	0.321	0.311	0.419	0.403	0.393	0.360	0.158	0.205	0.452	0.425	0.513	0.445	0.373	0.362
X_{24}	0.371	0.441	0.187	0.482	0.516	0.315	0.555	0.373	0.396	0.511	0.391	0.409	0.332	0.339	0.398
X_{25}	0.345	0.463	0.281	0.401	0.406	0.339	0.323	0.463	0.415	0.501	0.385	0.279	0.378	0.356	0.453
X_{26}	0.247	0.267	0.189	0.374	0.289	0.321	0.294	0.314	0.271	0.292	0.185	0.263	0.247	0.231	0.268

Continùa…

Sigue...

Matriz de Correlaciones (Datos Originales)
Tamaño de muestra n=209

Variables	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
X_{16}	1.000										
X_{17}	0.398	1.000									
X_{18}	0.289	0.587	1.000								
X_{19}	0.410	0.540	0.466	1.000							
X_{20}	0.431	0.485	0.472	0.539	1.000						
X_{21}	0.423	0.316	0.458	0.350	0.573	1.000					
X_{22}	0.417	0.319	0.421	0.293	0.452	0.668	1.000				
X_{23}	0.370	0.187	0.205	0.126	0.355	0.529	0.481	1.000			
X_{24}	0.400	0.427	0.314	0.412	0.554	0.410	0.322	0.400	1.000		
X_{25}	0.255	0.312	0.250	0.379	0.306	0.268	0.250	0.142	0.368	0.652	
X_{26}	0.358	0.290	0.228	0.265	0.255	0.215	0.171	0.195	0.332	0.177	0.995

Método de Eliminación por Filas

Matriz de Varianzas y Covarianzas (101 filas eliminadas)

Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	0.695														
X_2	0.453	0.800													
X_3	0.281	0.257	0.793												
X_4	0.478	0.400	0.311	0.830											
X_5	0.483	0.430	0.297	0.470	0.760										
X_6	0.363	0.337	0.196	0.416	0.357	0.672									
X_7	0.422	0.437	0.262	0.461	0.467	0.310	0.821								
X_8	0.499	0.466	0.317	0.362	0.486	0.264	0.435	1.348							
X_9	0.392	0.386	0.224	0.402	0.396	0.211	0.402	0.924	1.200						
X_{10}	0.473	0.483	0.302	0.608	0.473	0.409	0.454	0.430	0.384	0.741					
X_{11}	0.273	0.248	0.274	0.336	0.348	0.270	0.308	0.405	0.413	0.359	0.645				
X_{12}	0.341	0.347	0.178	0.386	0.337	0.348	0.311	0.286	0.343	0.388	0.306	0.534			
X_{13}	0.356	0.373	0.299	0.379	0.303	0.377	0.421	0.347	0.300	0.395	0.268	0.345	0.648		
X_{14}	0.442	0.480	0.259	0.410	0.500	0.374	0.429	0.418	0.398	0.482	0.414	0.394	0.381	0.970	
X_{15}	0.460	0.410	0.400	0.490	0.480	0.450	0.490	0.510	0.490	0.550	0.420	0.360	0.470	0.470	1.160
X_{16}	0.325	0.316	0.172	0.432	0.396	0.318	0.450	0.481	0.456	0.417	0.301	0.311	0.403	0.478	0.470
X_{17}	0.283	0.328	0.224	0.276	0.348	0.210	0.398	0.385	0.393	0.279	0.255	0.176	0.268	0.334	0.500
X_{18}	0.161	0.264	0.389	0.293	0.274	0.194	0.286	0.303	0.267	0.274	0.253	0.167	0.248	0.247	0.410
X_{19}	0.370	0.351	0.273	0.364	0.381	0.187	0.329	0.592	0.653	0.374	0.263	0.241	0.275	0.316	0.540
X_{20}	0.385	0.329	0.276	0.440	0.389	0.321	0.396	0.385	0.314	0.395	0.242	0.251	0.334	0.340	0.420
X_{21}	0.281	0.366	0.320	0.337	0.386	0.300	0.273	0.325	0.271	0.398	0.291	0.276	0.294	0.323	0.430
X_{22}	0.278	0.253	0.228	0.279	0.363	0.250	0.299	0.355	0.258	0.346	0.272	0.209	0.209	0.341	0.390
X_{23}	0.249	0.246	0.221	0.337	0.336	0.296	0.264	0.157	0.153	0.346	0.327	0.293	0.242	0.333	0.430
X_{24}	0.339	0.474	0.230	0.413	0.494	0.310	0.447	0.355	0.379	0.452	0.279	0.284	0.296	0.367	0.530
X_{25}	0.257	0.367	0.178	0.302	0.337	0.141	0.236	0.490	0.376	0.389	0.241	0.162	0.235	0.268	0.390
X_{26}	0.079	0.135	0.023	0.267	0.205	0.177	0.146	0.246	0.221	0.156	-0.008	0.125	0.141	0.013	0.190

Continúa…

Sigue...

Método de Eliminación por Filas
Matriz de Varianzas y Covarianzas (101 filas eliminadas)
Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
X_{16}	0.752										
X_{17}	0.291	0.725									
X_{18}	0.190	0.443	0.779								
X_{19}	0.344	0.443	0.391	0.908							
X_{20}	0.353	0.252	0.260	0.347	0.609						
X_{21}	0.298	0.161	0.278	0.285	0.359	0.595					
X_{22}	0.320	0.182	0.244	0.261	0.254	0.397	0.641				
X_{23}	0.280	0.137	0.111	0.099	0.240	0.322	0.316	0.659			
X_{24}	0.352	0.449	0.212	0.365	0.411	0.305	0.213	0.358	0.975		
X_{25}	0.273	0.271	0.236	0.476	0.285	0.282	0.264	0.164	0.418	0.741	
X_{26}	0.227	0.132	0.122	0.193	0.248	0.190	0.124	0.068	0.370	0.084	0.953

Método de Eliminación por Filas

Matriz de Correlaciones (101 filas eliminadas)

Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	1.000														
X_2	0.608	1.000													
X_3	0.378	0.323	1.000												
X_4	0.630	0.491	0.384	1.000											
X_5	0.665	0.551	0.383	0.592	1.000										
X_6	0.531	0.460	0.269	0.558	0.500	1.000									
X_7	0.559	0.540	0.325	0.559	0.592	0.417	1.000								
X_8	0.515	0.449	0.306	0.342	0.481	0.278	0.414	1.000							
X_9	0.429	0.394	0.230	0.403	0.415	0.235	0.405	0.727	1.000						
X_{10}	0.659	0.628	0.394	0.776	0.631	0.580	0.582	0.430	0.407	1.000					
X_{11}	0.408	0.346	0.383	0.459	0.497	0.411	0.424	0.434	0.469	0.519	1.000				
X_{12}	0.560	0.531	0.273	0.580	0.529	0.581	0.470	0.337	0.429	0.616	0.521	1.000			
X_{13}	0.531	0.518	0.418	0.516	0.432	0.571	0.577	0.371	0.340	0.570	0.414	0.586	1.000		
X_{14}	0.538	0.545	0.295	0.457	0.582	0.463	0.480	0.366	0.369	0.568	0.523	0.547	0.481	1.000	
X_{15}	0.512	0.426	0.417	0.499	0.511	0.510	0.502	0.408	0.415	0.593	0.485	0.457	0.542	0.443	1.000
X_{16}	0.449	0.407	0.222	0.547	0.523	0.448	0.572	0.478	0.480	0.558	0.433	0.490	0.577	0.559	0.503
X_{17}	0.399	0.431	0.295	0.356	0.469	0.301	0.516	0.389	0.421	0.381	0.373	0.283	0.390	0.398	0.545
X_{18}	0.219	0.334	0.495	0.365	0.356	0.268	0.357	0.296	0.276	0.360	0.357	0.259	0.349	0.284	0.431
X_{19}	0.465	0.412	0.322	0.420	0.459	0.239	0.381	0.535	0.625	0.456	0.343	0.346	0.359	0.336	0.526
X_{20}	0.591	0.472	0.397	0.619	0.572	0.501	0.560	0.425	0.367	0.588	0.387	0.440	0.532	0.442	0.500
X_{21}	0.437	0.531	0.466	0.480	0.574	0.474	0.391	0.363	0.321	0.599	0.470	0.490	0.474	0.425	0.518
X_{22}	0.416	0.353	0.320	0.382	0.520	0.381	0.413	0.382	0.294	0.503	0.423	0.357	0.325	0.433	0.452
X_{23}	0.369	0.339	0.306	0.456	0.476	0.445	0.360	0.167	0.172	0.495	0.501	0.493	0.371	0.417	0.492
X_{24}	0.412	0.537	0.262	0.459	0.574	0.384	0.500	0.309	0.350	0.532	0.352	0.393	0.373	0.377	0.498
X_{25}	0.358	0.476	0.232	0.385	0.449	0.199	0.303	0.490	0.399	0.524	0.348	0.258	0.339	0.316	0.421
X_{26}	0.097	0.155	0.027	0.301	0.241	0.222	0.165	0.217	0.207	0.186	-0.010	0.175	0.180	0.013	0.181

Continúa…

Sigue...

Método de Eliminación por Filas
Matriz de Correlaciones (101 filas eliminadas)
Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
X_{16}	1.000										
X_{17}	0.395	1.000									
X_{18}	0.249	0.590	1.000								
X_{19}	0.416	0.546	0.464	1.000							
X_{20}	0.522	0.380	0.378	0.466	1.000						
X_{21}	0.445	0.245	0.409	0.388	0.597	1.000					
X_{22}	0.462	0.267	0.346	0.342	0.407	0.643	1.000				
X_{23}	0.397	0.198	0.156	0.129	0.378	0.514	0.486	1.000			
X_{24}	0.412	0.534	0.243	0.388	0.533	0.401	0.269	0.447	1.000		
X_{25}	0.366	0.369	0.311	0.580	0.424	0.425	0.382	0.234	0.491	1.000	
X_{26}	0.268	0.159	0.142	0.208	0.325	0.253	0.159	0.085	0.384	0.100	1.000

Se puede apreciar en la matriz de varianzas y covarianzas de la matriz de datos originales, muestra que la mayor covarianza se da entre las variables "Saben trabajar en equipo" y "*Saben desarrollar actividades conjuntas con profesionales de áreas diferentes a la suya",* esto es 0.749 y la menor covarianza es entre las variables "*Sólida formación en Informática"* y "*Muestran clara sensibilidad Social y Humana".*

Así como también se puede notar que la más alta correlación con la variable *Retos*, se presenta con la variable *Superación*, la misma que alcanza un valor de 0.664, por el contrario, para las proposiciones "*Trabajar en Equipo"* y *"Conocimientos de Informática"*, los coeficientes de correlación son muy cercanos a cero por lo que se concluye que no existe relación lineal entre estas variables y el "*Formación comparable con extranjeros"*

La correlación más fuerte se presenta entre las variables "Saben trabajar en equipo" y "*Saben desarrollar actividades conjuntas con profesionales de áreas diferentes a la suya"*, esta correlación es de 0.725, seguida por la correlación entre las variables "*Son altamente capacitados para llevar a cabo Análisis Cuantitativos"* y *"Tienen alta compresión de los principios Físicos y Naturales"* (0.668).

Mientras que en la matriz de varianzas y covarianzas con filas eliminadas, la covarianza entre "Saben trabajar en equipo" y "*Saben*

desarrollar actividades conjuntas con profesionales de áreas diferentes a la suya", aumenta su valor es decir de 0.749 a 0.924, y ahora la menor covarianza se da entre las variables "*Son personas con capacidad de análisis para llegar a conclusiones válidas, bajo distintas circunstancias"* y "*Poseen el nivel de Inglés adecuado para utilizarlo de la manera requerida por sus actividades en la Organización",* la correlación entre las varibles tambièn cambia en la matriz de correlaciones con filas eliminadas, ya que ahora la mayor correlación es entre las variables "*Tienen capacidad para manejar los retos e innovaciones" y "Tienen claros propósitos de superación, esto es: tenacidad y estrategia"* (0.776), y la menor correlación se da entre "*Son personas con capacidad de análisis para llegar a conclusiones válidas, bajo distintas circunstancias"* y "*Poseen el nivel de Inglés adecuado para utilizarlo de la manera requerida por sus actividades en la Organización"*(0.097).

En el siguiente Cuadro podemos apreciar los estimadores para las variables que tienen datos faltantes, donde la media en la variable "Formación Comparable" con 26% de datos eliminados aumenta de 3.938 a 3.990, así como también su varianza. Mientras que en la variable "Proceso de Ascenso", la varianza disminuye de 0.812 a 0.752 y el valor de la media aumenta de 3.770 a 3.780

Método de Eliminación por Filas

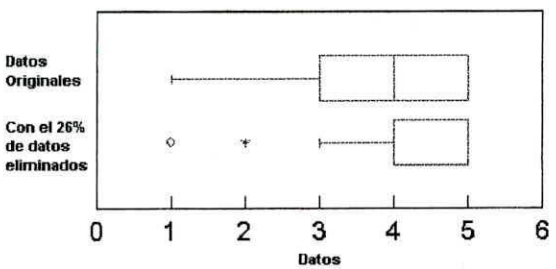
Tamaño de muestra n=209 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Formación Comparable*" y "*Proceso de Ascenso*"

Estimadores "Formación Comparable "

Estimadores		Datos Originales	Con el 26% de datos eliminadas
n		209	101
Media		3,938	3,990
Mediana		4,000	4,000
Moda		4,000	4,000
Varianza		0,943	0,970
Desviación Estándar		0,971	0,984
Error Estándar		0.067	0.098
Coeficiente de Asimetría		-0,924	-1,069
Curtosis		0,655	1,073
Rango		4,000	4,000
Mínimo		1,000	1,000
Máximo		5,000	5,000
Percentiles	25	3,000	4,000
	50	4,000	4,000
	75	5,000	5,000

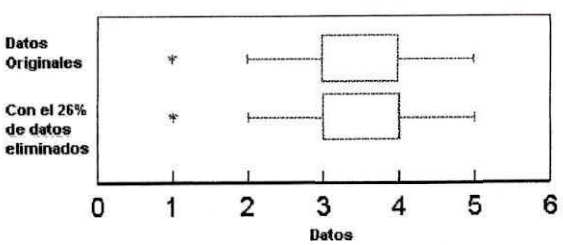
Diagrama de Cajas "Formación Comparable "



Estimadores "Proceso de Ascenso"

Estimadores		Datos Originales	Con el 26% de datos eliminadas
n		209	101
Media		3,770	3,782
Mediana		4,000	4,000
Moda		4,000	4,000
Varianza		0,812	0,752
Desviación Estándar		0,901	0,867
Error Estándar		0.062	0.086
Coeficiente de Asimetría		-0,763	-0,685
Curtosis		0,797	0,441
Rango		4,000	4,000
Mínimo		1,000	1,000
Máximo		5,000	5,000
Percentiles	25	3,000	3,000
	50	4,000	4,000
	75	4,000	4,000

Diagrama de Cajas "Proceso de Ascenso"



Método de Imputación por la Media y Regresión

Estos métodos se aplican a la misma matriz de datos utilizada en el método de eliminación por filas, es decir se completan datos en las variables "Su *formación es comparable a la de profesionales extranjeros"* y "*Su proceso de ascenso en el organigrama de la organización es notable"*, que presentan cincuenta y cuatro valores faltantes cada una. A través del Método de Imputación por Media, se procede a calcular la media aritmética de la variable "Su *formación es comparable a la de profesionales extranjeros"* con los cincuenta y cuatro datos faltantes, cuyo valor es 1.928, así como también la media de la variable "*Su proceso de ascenso en el organigrama de la organización es notable",* 1.828; estos valores se reemplazan en los datos faltantes de cada variable.

En las siguientes tablas se realiza una comparación entre el dato observado y el valor con *imputación por la media y regresión*, donde se puede notar que la diferencia en valor absoluto entre el dato observado y el estimado de cada variable es menor en el "Método de Imputación por Regresión".

Comparación de los Métodos de Imputación

Tamaño de muestra n=209 y 5% de datos faltantes en la matriz

Datos completados en "Su formación es comparable" por la Media

Dato Observado	Resultado Imputación por Media	Error | Dato Observado – Resultado de Imputación por Media |
5	3.974	1,026
5	3.974	1,026
5	3.974	1,026
4	3.974	0,026
3	3.974	0,974
4	3.974	0,026
3	3.974	0,974
5	3.974	1,026
3	3.974	0,974
5	3.974	1,026
4	3.974	0,026
2	3.974	1,974
4	3.974	0,026
5	3.974	1,026
3	3.974	0,974
3	3.974	0,974
4	3.974	0,026
4	3.974	0,026
2	3.974	1,974
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026
1	3.974	2,974
4	3.974	0,026
3	3.974	0,974
5	3.974	1,026
5	3.974	1,026
4	3.974	0,026
5	3.974	1,026
1	3.974	2,974
5	3.974	1,026
4	3.974	0,026
3	3.974	0,974
5	3.974	1,026
4	3.974	0,026
3	3.974	0,974
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026
5	3.974	1,026
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026
3	3.974	0,974
2	3.974	1,974
3	3.974	0,974
5	3.974	1,026
4	3.974	0,026
4	3.974	0,026
4	3.974	0,026

Datos completados en "Su formación es comparable" por Regresión

Dato Observado	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
5	4.976	0,024
5	4.947	0,053
5	4.985	0,015
4	4.023	0,023
3	2.837	0,163
4	3.989	0,011
3	3.165	0,165
5	4.879	0,121
3	3.101	0,101
5	5.392	0,392
4	4.083	0,083
2	2.221	0,221
4	4.475	0,475
5	4.789	0,211
3	3.058	0,058
3	2.882	0,118
4	3.809	0,191
4	3.996	0,004
2	2.145	0,145
4	4.165	0,165
4	4.085	0,085
4	3.982	0,018
4	3.993	0,007
4	3.991	0,009
1	0.972	0,028
4	3.983	0,017
3	2.995	0,005
5	4.863	0,137
5	4.947	0,053
4	3.981	0,019
5	5.018	0,018
1	1.103	0,103
5	4.972	0,028
4	3.981	0,019
3	2.993	0,007
5	4.971	0,029
4	3.986	0,014
3	2.975	0,025
4	3.991	0,009
4	3.883	0,117
4	3.980	0,020
5	5.005	0,005
4	3.992	0,008
4	4.001	0,001
4	4.085	0,085
4	4.101	0,101
4	4.003	0,003
3	3.103	0,103
2	1.992	0,008
3	2.933	0,067
5	4.972	0,028
4	3.993	0,007
4	3.985	0,015
4	3.983	0,017

Comparación de los Métodos de Imputación

Tamaño de muestra n=209 y 5% de datos faltantes en la matriz

Datos completados en "Proceso de Ascenso" por la Media

Dato Observado	Resultado Imputación por Media	Error | Dato Observado – Resultado de Imputación por Media |
5	3.806	1,194
3	3.806	0,806
3	3.806	0,806
3	3.806	0,806
4	3.806	0,194
4	3.806	0,194
4	3.806	0,194
5	3.806	1,194
4	3.806	0,194
5	3.806	1,194
1	3.806	2,806
5	3.806	1,194
4	3.806	0,194
3	3.806	0,806
5	3.806	1,194
4	3.806	0,194
2	3.806	1,806
3	3.806	0,806
1	3.806	2,806
4	3.806	0,194
4	3.806	0,194
3	3.806	0,806
4	3.806	0,194
3	3.806	0,806
4	3.806	0,194
4	3.806	0,194
4	3.806	0,194
4	3.806	0,194
5	3.806	1,194
4	3.806	0,194
5	3.806	1,194
2	3.806	1,806
4	3.806	0,194
4	3.806	0,194
5	3.806	1,194
4	3.806	0,194
3	3.806	0,806
3	3.806	0,806
4	3.806	0,194
3	3.806	0,806
5	3.806	1,194
4	3.806	0,194
3	3.806	0,806
2	3.806	1,806
3	3.806	0,806
3	3.806	0,806
5	3.806	1,194
3	3.806	0,806
5	3.806	1,194
3	3.806	0,806
4	3.806	0,194
3	3.806	0,806
4	3.806	0,194
3	3.806	0,806

Datos completados en "Proceso de Ascenso" por Regresión

Dato Observado	Resultado de Predicción	Error | Dato Observado – Resultado de Predicción |
5	4.981	0,019
3	3.101	0,101
3	3.054	0,054
3	3.082	0,082
4	4.032	0,032
4	4.101	0,101
4	4.003	0,003
5	4.999	0,001
4	3.972	0,028
5	5.004	0,004
1	0.987	0,013
5	4.972	0,028
4	4.009	0,009
3	2.898	0,102
5	4.932	0,068
4	3.901	0,099
2	2.005	0,005
3	2.992	0,008
1	1.083	0,083
4	3.983	0,017
4	3.972	0,028
3	2.995	0,005
4	4.015	0,015
3	3.200	0,200
4	4.108	0,108
4	3.983	0,017
4	3.974	0,026
4	4.073	0,073
5	4.985	0,015
4	3.932	0,068
5	4.871	0,129
2	1.993	0,007
4	3.982	0,018
4	3.991	0,009
5	4.993	0,007
4	3.932	0,068
3	2.971	0,029
3	2.992	0,008
4	3.931	0,069
3	2.898	0,102
5	4.901	0,099
4	3.907	0,093
3	2.911	0,089
2	1.906	0,094
3	3.072	0,072
3	3.031	0,031
5	4.931	0,069
3	2.922	0,078
5	4.972	0,028
3	2.983	0,017
4	3.915	0,085
3	2.909	0,091
4	3.933	0,067
3	3.081	0,081

Método de Imputación por la Media y Regresión

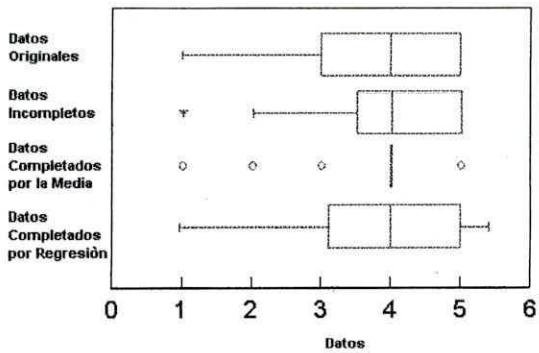
Tamaño de muestra n=209 y 5% de datos faltantes en la matriz

Tabla y Diagrama de la "*Su formación es comparable*" y "*Proceso de Ascenso*"

Estimadores "*Su formación es comparable*"

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		209	155	209	209
Media		3,938	3,974	3,974	3,941
Mediana		4,000	4,000	4,000	4,000
Moda		4,000	4,000	4,000	4,000
Varianza		0,943	0,934	0,692	0,935
Desviación Estándar		0,971	0,967	0,832	0,967
Error Estándar		0.067	0.078	0.058	0.067
Coeficiente de Asimetría		-0,924	-0,909	-1,053	-0,918
Curtosis		0,655	0,548	1,780	0,648
Rango		4,000	4,000	4,000	4,420
Mínimo		1,000	1,000	1,000	0,970
Máximo		5,000	5,000	5,000	5,390
Percentiles	25	3,000	3,000	3,974	3,102
	50	4,000	4,000	4,000	4,000
	75	5,000	5,000	4,000	5,000

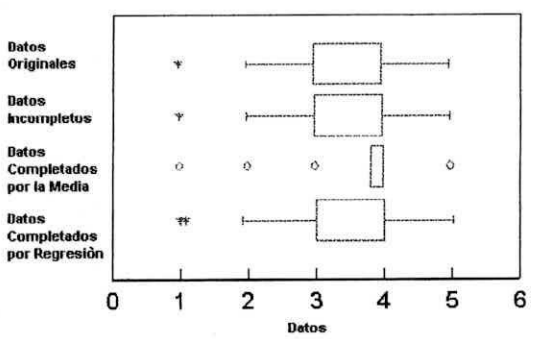
Diagrama de Cajas "*Su formación es comparable*"



Estimadores "*Proceso de Ascenso*"

Estimadores		Datos Originales	Datos Incompletos	Datos Completados por la Media	Datos Completados por Regresión
n		209	155	209	209
Media		3,770	3,807	3,806	3,767
Mediana		4,000	4,000	4,000	4,000
Moda		4,000	4,000	4,000	4,000
Varianza		0,812	0,755	0,559	0,806
Desviación Estándar		0,901	0,869	0,747	0,898
Error Estándar		0.062	0.069	0.052	0.062
Coeficiente de Asimetría		-0,763	-0,817	-0,946	-0,769
Curtosis		0,797	1,028	2,421	0,794
Rango		4,000	4,000	4,000	4,020
Mínimo		1,000	1,000	1,000	0,990
Máximo		5,000	5,000	5,000	5,000
Percentiles	25	3,000	3,000	3,806	3,000
	50	4,000	4,000	4,000	4,000
	75	4,000	4,000	4,000	4,000

Diagrama de Cajas "*Proceso de Ascenso*"



El vector de medias con 271 datos en total completados por la media en "Su *formación es comparable a la de profesionales extranjeros"* y en "*Su proceso de ascenso en el organigrama de la organización es notable"*, es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \\ \overline{X}_7 \\ \overline{X}_8 \\ \overline{X}_9 \\ \overline{X}_{10} \\ \overline{X}_{11} \\ \overline{X}_{12} \\ \overline{X}_{13} \\ \overline{X}_{14} \\ \overline{X}_{15} \\ \overline{X}_{16} \\ \overline{X}_{17} \\ \overline{X}_{18} \\ \overline{X}_{19} \\ \overline{X}_{20} \\ \overline{X}_{21} \\ \overline{X}_{22} \\ \overline{X}_{23} \\ \overline{X}_{24} \end{pmatrix} = \begin{pmatrix} 4.091 \\ 3.962 \\ 3.746 \\ 4.005 \\ 3.943 \\ 4.278 \\ 3.789 \\ 3.852 \\ 3.861 \\ 4.124 \\ 4.378 \\ 4.311 \\ 4.139 \\ 3.974 \\ 3.938 \\ 3.806 \\ 3.431 \\ 3.699 \\ 3.560 \\ 3.952 \\ 4.225 \\ 4.134 \\ 4.254 \\ 3.584 \\ 3.856 \\ 3.196 \end{pmatrix}$$

Mientras que el vector de medias con 271 datos completados por la regresión en "Su *formación es comparable a la de profesionales*

extranjeros" y en "*Su proceso de ascenso en el organigrama de la organización es notable"* es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \\ \overline{X}_5 \\ \overline{X}_6 \\ \overline{X}_7 \\ \overline{X}_8 \\ \overline{X}_9 \\ \overline{X}_{10} \\ \overline{X}_{11} \\ \overline{X}_{12} \\ \overline{X}_{13} \\ \overline{X}_{14} \\ \overline{X}_{15} \\ \overline{X}_{16} \\ \overline{X}_{17} \\ \overline{X}_{18} \\ \overline{X}_{19} \\ \overline{X}_{20} \\ \overline{X}_{21} \\ \overline{X}_{22} \\ \overline{X}_{23} \\ \overline{X}_{24} \end{pmatrix} = \begin{pmatrix} 4.091 \\ 3.962 \\ 3.746 \\ 4.005 \\ 3.943 \\ 4.278 \\ 3.789 \\ 3.852 \\ 3.861 \\ 4.124 \\ 4.378 \\ 4.311 \\ 4.139 \\ 3.941 \\ 3.938 \\ 3.767 \\ 3.431 \\ 3.699 \\ 3.560 \\ 3.952 \\ 4.225 \\ 4.134 \\ 4.254 \\ 3.584 \\ 3.856 \\ 3.196 \end{pmatrix}$$

El efecto que causa en la *matriz de varianzas y covarianzas* y *matriz de correlaciones*, el completar 5% de datos faltantes en una matriz de tamaño 209, por medio de la imputación por media y regresión, se presenta en las siguientes ocho páginas.

Método de Imputación por Media

Matriz de Varianzas y Covarianzas

Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	0.679														
X_2	0.441	0.739													
X_3	0.312	0.322	1.007												
X_4	0.432	0.414	0.328	0.880											
X_5	0.419	0.402	0.283	0.481	0.679										
X_6	0.374	0.352	0.253	0.427	0.338	0.721									
X_7	0.399	0.372	0.268	0.515	0.435	0.333	0.773								
X_8	0.403	0.403	0.318	0.414	0.352	0.282	0.416	1.127							
X_9	0.311	0.355	0.229	0.380	0.333	0.250	0.350	0.749	0.947						
X_{10}	0.364	0.404	0.238	0.523	0.430	0.384	0.406	0.379	0.388	0.706					
X_{11}	0.235	0.231	0.231	0.354	0.286	0.298	0.282	0.388	0.332	0.309	0.630				
X_{12}	0.303	0.315	0.204	0.450	0.345	0.375	0.345	0.282	0.289	0.365	0.339	0.600			
X_{13}	0.329	0.337	0.338	0.374	0.301	0.413	0.366	0.309	0.269	0.353	0.269	0.385	0.649		
X_{14}	0.296	0.273	0.170	0.256	0.316	0.256	0.276	0.214	0.194	0.243	0.282	0.271	0.253	0.692	
X_{15}	0.376	0.401	0.388	0.390	0.347	0.416	0.357	0.414	0.381	0.460	0.341	0.332	0.388	0.336	1.001
X_{16}	0.283	0.286	0.162	0.393	0.310	0.281	0.350	0.344	0.325	0.297	0.223	0.275	0.279	0.230	0.288
X_{17}	0.278	0.305	0.244	0.363	0.323	0.279	0.370	0.338	0.373	0.321	0.245	0.197	0.252	0.229	0.431
X_{18}	0.249	0.296	0.390	0.357	0.295	0.204	0.326	0.301	0.285	0.297	0.268	0.219	0.278	0.205	0.375
X_{19}	0.266	0.300	0.215	0.387	0.330	0.214	0.344	0.482	0.516	0.339	0.259	0.234	0.234	0.188	0.386
X_{20}	0.312	0.296	0.243	0.462	0.358	0.311	0.442	0.339	0.320	0.381	0.259	0.303	0.305	0.259	0.372
X_{21}	0.297	0.364	0.331	0.364	0.335	0.336	0.288	0.255	0.243	0.357	0.299	0.324	0.353	0.231	0.375
X_{22}	0.257	0.274	0.260	0.302	0.291	0.270	0.245	0.241	0.235	0.277	0.238	0.256	0.255	0.243	0.326
X_{23}	0.241	0.236	0.267	0.335	0.284	0.285	0.270	0.144	0.170	0.324	0.288	0.339	0.306	0.212	0.309
X_{24}	0.298	0.369	0.182	0.440	0.413	0.260	0.474	0.385	0.375	0.417	0.302	0.308	0.260	0.252	0.387
X_{25}	0.229	0.321	0.228	0.304	0.271	0.232	0.229	0.397	0.326	0.340	0.247	0.175	0.246	0.190	0.366
X_{26}	0.203	0.229	0.189	0.350	0.237	0.272	0.258	0.332	0.263	0.245	0.147	0.203	0.199	0.066	0.267

Continúa…

Sigue...

Método de Imputación por Media
Matriz de Varianzas y Covarianzas
Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
X_{16}	0.559										
X_{17}	0.211	0.823									
X_{18}	0.165	0.467	0.769								
X_{19}	0.241	0.445	0.371	0.825							
X_{20}	0.244	0.352	0.332	0.392	0.642						
X_{21}	0.230	0.220	0.308	0.244	0.352	0.589					
X_{22}	0.232	0.216	0.276	0.199	0.271	0.383	0.559				
X_{23}	0.242	0.145	0.154	0.098	0.243	0.347	0.307	0.729			
X_{24}	0.281	0.377	0.268	0.364	0.432	0.306	0.234	0.332	0.946		
X_{25}	0.178	0.312	0.250	0.379	0.306	0.268	0.250	0.142	0.368	0.652	
X_{26}	0.244	0.290	0.228	0.265	0.255	0.215	0.171	0.195	0.332	0.177	0.995

Método de Imputación por Media

Matriz de Correlaciones

Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	1.000														
X_2	0.623	1.000													
X_3	0.377	0.373	1.000												
X_4	0.559	0.513	0.349	1.000											
X_5	0.616	0.567	0.343	0.622	1.000										
X_6	0.534	0.482	0.297	0.536	0.483	1.000									
X_7	0.551	0.492	0.304	0.625	0.600	0.446	1.000								
X_8	0.461	0.442	0.298	0.416	0.402	0.313	0.445	1.000							
X_9	0.387	0.425	0.235	0.417	0.416	0.303	0.410	0.725	1.000						
X_{10}	0.525	0.559	0.283	0.664	0.621	0.538	0.550	0.425	0.474	1.000					
X_{11}	0.359	0.338	0.290	0.475	0.437	0.443	0.404	0.460	0.429	0.463	1.000				
X_{12}	0.475	0.473	0.263	0.620	0.540	0.570	0.506	0.343	0.383	0.561	0.551	1.000			
X_{13}	0.495	0.487	0.418	0.495	0.454	0.604	0.517	0.362	0.344	0.521	0.421	0.616	1.000		
X_{14}	0.431	0.382	0.204	0.328	0.461	0.363	0.377	0.242	0.240	0.347	0.427	0.421	0.377	1.000	
X_{15}	0.456	0.467	0.386	0.415	0.421	0.490	0.406	0.390	0.391	0.547	0.429	0.428	0.482	0.403	1.000
X_{16}	0.460	0.445	0.216	0.560	0.503	0.443	0.533	0.434	0.446	0.474	0.376	0.476	0.464	0.369	0.385
X_{17}	0.372	0.391	0.268	0.427	0.432	0.362	0.464	0.351	0.422	0.421	0.340	0.281	0.345	0.304	0.475
X_{18}	0.344	0.393	0.442	0.434	0.408	0.274	0.422	0.324	0.334	0.404	0.385	0.323	0.393	0.281	0.428
X_{19}	0.356	0.385	0.236	0.454	0.441	0.278	0.431	0.501	0.583	0.444	0.359	0.332	0.321	0.249	0.425
X_{20}	0.473	0.430	0.302	0.614	0.542	0.458	0.627	0.399	0.411	0.566	0.406	0.489	0.472	0.388	0.464
X_{21}	0.469	0.553	0.430	0.506	0.530	0.516	0.427	0.313	0.325	0.553	0.491	0.545	0.572	0.362	0.488
X_{22}	0.417	0.427	0.347	0.431	0.473	0.426	0.372	0.304	0.323	0.440	0.400	0.442	0.424	0.392	0.435
X_{23}	0.343	0.321	0.311	0.419	0.403	0.393	0.360	0.158	0.205	0.452	0.425	0.513	0.445	0.299	0.362
X_{24}	0.371	0.441	0.187	0.482	0.516	0.315	0.555	0.373	0.396	0.511	0.391	0.409	0.332	0.312	0.398
X_{25}	0.345	0.463	0.281	0.401	0.406	0.339	0.323	0.463	0.415	0.501	0.385	0.279	0.378	0.282	0.453
X_{26}	0.247	0.267	0.189	0.374	0.289	0.321	0.294	0.314	0.271	0.292	0.185	0.263	0.247	0.080	0.268

Continúa…

Sigue...

Método de Imputación por Media
Matriz de Correlaciones
Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
X_{16}	1.000										
X_{17}	0.312	1.000									
X_{18}	0.251	0.587	1.000								
X_{19}	0.356	0.540	0.466	1.000							
X_{20}	0.408	0.485	0.472	0.539	1.000						
X_{21}	0.401	0.316	0.458	0.350	0.573	1.000					
X_{22}	0.415	0.319	0.421	0.293	0.452	0.668	1.000				
X_{23}	0.379	0.187	0.205	0.126	0.355	0.529	0.481	1.000			
X_{24}	0.386	0.427	0.314	0.412	0.554	0.410	0.322	0.400	1.000		
X_{25}	0.294	0.426	0.353	0.517	0.472	0.433	0.414	0.206	0.468	1.000	
X_{26}	0.327	0.321	0.260	0.292	0.319	0.281	0.229	0.229	0.342	0.220	1.000

Método de Imputación por Regresión

Matriz de Varianzas y Covarianzas

Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	0.679														
X_2	0.441	0.739													
X_3	0.312	0.322	1.007												
X_4	0.432	0.414	0.328	0.880											
X_5	0.419	0.402	0.283	0.481	0.679										
X_6	0.374	0.352	0.253	0.427	0.338	0.721									
X_7	0.399	0.372	0.268	0.515	0.435	0.333	0.773								
X_8	0.403	0.403	0.318	0.414	0.352	0.282	0.416	1.127							
X_9	0.311	0.355	0.229	0.380	0.333	0.250	0.350	0.749	0.947						
X_{10}	0.364	0.404	0.238	0.523	0.430	0.384	0.406	0.379	0.388	0.706					
X_{11}	0.235	0.231	0.231	0.354	0.286	0.298	0.282	0.388	0.332	0.309	0.630				
X_{12}	0.303	0.315	0.204	0.450	0.345	0.375	0.345	0.282	0.289	0.365	0.339	0.600			
X_{13}	0.329	0.337	0.338	0.374	0.301	0.413	0.366	0.309	0.269	0.353	0.269	0.385	0.649		
X_{14}	0.434	0.413	0.364	0.417	0.418	0.411	0.375	0.357	0.309	0.351	0.373	0.392	0.372	0.935	
X_{15}	0.376	0.401	0.388	0.390	0.347	0.416	0.357	0.414	0.381	0.460	0.341	0.332	0.388	0.447	1.001
X_{16}	0.360	0.350	0.200	0.453	0.360	0.324	0.387	0.384	0.357	0.358	0.255	0.331	0.342	0.409	0.391
X_{17}	0.278	0.305	0.244	0.363	0.323	0.279	0.370	0.338	0.373	0.321	0.245	0.197	0.252	0.319	0.431
X_{18}	0.249	0.296	0.390	0.357	0.295	0.204	0.326	0.301	0.285	0.297	0.268	0.219	0.278	0.344	0.375
X_{19}	0.266	0.300	0.215	0.387	0.330	0.214	0.344	0.482	0.516	0.339	0.259	0.234	0.234	0.274	0.386
X_{20}	0.312	0.296	0.243	0.462	0.358	0.311	0.442	0.339	0.320	0.381	0.259	0.303	0.305	0.345	0.372
X_{21}	0.297	0.364	0.331	0.364	0.335	0.336	0.288	0.255	0.243	0.357	0.299	0.324	0.353	0.349	0.375
X_{22}	0.257	0.274	0.260	0.302	0.291	0.270	0.245	0.241	0.235	0.277	0.238	0.256	0.255	0.334	0.326
X_{23}	0.241	0.236	0.267	0.335	0.284	0.285	0.270	0.144	0.170	0.324	0.288	0.339	0.306	0.306	0.309
X_{24}	0.298	0.369	0.182	0.440	0.413	0.260	0.474	0.385	0.375	0.417	0.302	0.308	0.260	0.321	0.387
X_{25}	0.229	0.321	0.228	0.304	0.271	0.232	0.229	0.397	0.326	0.340	0.247	0.175	0.246	0.281	0.366
X_{26}	0.203	0.229	0.189	0.350	0.237	0.272	0.258	0.332	0.263	0.245	0.147	0.203	0.199	0.219	0.267

Continúa…

Sigue...

Método de Imputación por Regresión
Matriz de Varianzas y Covarianzas
Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
X_{16}	0.806										
X_{17}	0.326	0.823									
X_{18}	0.228	0.467	0.769								
X_{19}	0.338	0.445	0.371	0.825							
X_{20}	0.311	0.352	0.332	0.392	0.642						
X_{21}	0.290	0.220	0.308	0.244	0.352	0.589					
X_{22}	0.278	0.216	0.276	0.199	0.271	0.383	0.559				
X_{23}	0.283	0.145	0.154	0.098	0.243	0.347	0.307	0.729			
X_{24}	0.349	0.377	0.268	0.364	0.432	0.306	0.234	0.332	0.946		
X_{25}	0.254	0.312	0.250	0.379	0.306	0.268	0.250	0.142	0.368	0.652	
X_{26}	0.358	0.290	0.228	0.265	0.255	0.215	0.171	0.195	0.332	0.177	0.995

Método de Imputación por Regresión

Matriz de Correlaciones

Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	1.000														
X_2	0.623	1.000													
X_3	0.377	0.373	1.000												
X_4	0.559	0.513	0.349	1.000											
X_5	0.616	0.567	0.343	0.622	1.000										
X_6	0.534	0.482	0.297	0.536	0.483	1.000									
X_7	0.551	0.492	0.304	0.625	0.600	0.446	1.000								
X_8	0.461	0.442	0.298	0.416	0.402	0.313	0.445	1.000							
X_9	0.387	0.425	0.235	0.417	0.416	0.303	0.410	0.725	1.000						
X_{10}	0.525	0.559	0.283	0.664	0.621	0.538	0.550	0.425	0.474	1.000					
X_{11}	0.359	0.338	0.290	0.475	0.437	0.443	0.404	0.460	0.429	0.463	1.000				
X_{12}	0.475	0.473	0.263	0.620	0.540	0.570	0.506	0.343	0.383	0.561	0.551	1.000			
X_{13}	0.495	0.487	0.418	0.495	0.454	0.604	0.517	0.362	0.344	0.521	0.421	0.616	1.000		
X_{14}	0.544	0.496	0.375	0.459	0.524	0.501	0.441	0.348	0.328	0.433	0.486	0.523	0.477	1.000	
X_{15}	0.456	0.467	0.386	0.415	0.421	0.490	0.406	0.390	0.391	0.547	0.429	0.428	0.482	0.462	1.000
X_{16}	0.486	0.454	0.221	0.538	0.487	0.426	0.491	0.403	0.409	0.475	0.358	0.476	0.473	0.471	0.435
X_{17}	0.372	0.391	0.268	0.427	0.432	0.362	0.464	0.351	0.422	0.421	0.340	0.281	0.345	0.363	0.475
X_{18}	0.344	0.393	0.442	0.434	0.408	0.274	0.422	0.324	0.334	0.404	0.385	0.323	0.393	0.405	0.428
X_{19}	0.356	0.385	0.236	0.454	0.441	0.278	0.431	0.501	0.583	0.444	0.359	0.332	0.321	0.312	0.425
X_{20}	0.473	0.430	0.302	0.614	0.542	0.458	0.627	0.399	0.411	0.566	0.406	0.489	0.472	0.445	0.464
X_{21}	0.469	0.553	0.430	0.506	0.530	0.516	0.427	0.313	0.325	0.553	0.491	0.545	0.572	0.471	0.488
X_{22}	0.417	0.427	0.347	0.431	0.473	0.426	0.372	0.304	0.323	0.440	0.400	0.442	0.424	0.462	0.435
X_{23}	0.343	0.321	0.311	0.419	0.403	0.393	0.360	0.158	0.205	0.452	0.425	0.513	0.445	0.371	0.362
X_{24}	0.371	0.441	0.187	0.482	0.516	0.315	0.555	0.373	0.396	0.511	0.391	0.409	0.332	0.342	0.398
X_{25}	0.345	0.463	0.281	0.401	0.406	0.339	0.323	0.463	0.415	0.501	0.385	0.279	0.378	0.360	0.453
X_{26}	0.247	0.267	0.189	0.374	0.289	0.321	0.294	0.314	0.271	0.292	0.185	0.263	0.247	0.227	0.268

Continúa…

Sigue...

Método de Imputación por Regresión
Matriz de Correlaciones
Tamaño de muestra n=209, 5% de datos faltantes en la matriz

Variables	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
X_{16}	1.000										
X_{17}	0.400	1.000									
X_{18}	0.290	0.587	1.000								
X_{19}	0.414	0.540	0.466	1.000							
X_{20}	0.432	0.485	0.472	0.539	1.000						
X_{21}	0.421	0.316	0.458	0.350	0.573	1.000					
X_{22}	0.414	0.319	0.421	0.293	0.452	0.668	1.000				
X_{23}	0.369	0.187	0.205	0.126	0.355	0.529	0.481	1.000			
X_{24}	0.400	0.427	0.314	0.412	0.554	0.410	0.322	0.400	1.000		
X_{25}	0.350	0.426	0.353	0.517	0.472	0.433	0.414	0.206	0.468	1.000	
X_{26}	0.400	0.321	0.260	0.292	0.319	0.281	0.229	0.229	0.342	0.220	1.000

En la matriz de varianzas y covarianzas de los datos completados por el Método de Imputación por media, podemos apreciar que las únicas covarianzas que cambian, son las de las variables a las cuales se les completó datos, donde la mayoría de las covarianzas disminuyen; tal es el caso de la covarianza entre "*Son personas con capacidad de análisis para llegar a conclusiones válidas, bajo distintas circunstancias"* y "Su *formación es comparable a la de profesionales Extranjeros",* que disminuye de 0.442 (valor de la matriz de datos originales) a 0.296, así como también la covarianza entre "*Tienen capacidad para manejar los retos e innovaciones"* y "*Su proceso de ascenso en el organigrama de la Organización es notable"* , que disminuye de 0.432 a 0.393.

En la matriz de correlaciones de los datos completados por medio de Imputación por media se aprecia que, la correlación entre "*Saben desarrollar actividades conjuntas con profesionales de áreas diferentes a la suya"* y "Su *formación es comparable a la de profesionales Extranjeros"* disminuye de 0.369(en la matriz de correlación de datos originales) a 0.240, mientras que la correlación entre "Su *formación es comparable a la de profesionales Extranjeros"* y "*Poseen el nivel de Inglés adecuado para utilizarlo de la manera requerida por sus actividades en la Organización"* aumenta de 0.013 a 0.080.

Por otro lado, en la matriz de varianzas y covarianzas de de los datos completados utilizando regresión, la covarianza entre "*Son personas con capacidad de análisis para llegar a conclusiones válidas, bajo distintas circunstancias*" y "Su *formación es comparable a la de profesionales Extranjeros*", es de 0.434 (covarianza que tiende al verdadero valor de la matriz de varianzas y covarianzas de los datos originales, 0.442), así como también la covarianza entre "*Tienen capacidad para manejar los retos e innovaciones*" y "*Su proceso de ascenso en el organigrama de la Organización es notable*", que es de 0.453.

En la matriz de correlaciones de los datos completados por regresión, la correlación entre "*Saben desarrollar actividades conjuntas con profesionales de áreas diferentes a la suya*" y "Su *formación es comparable a la de profesionales Extranjeros*" disminuye de 0.369(en la matriz de correlación de datos originales) a 0.329, mientras que la correlación entre "Su *formación es comparable a la de profesionales Extranjeros*" y "*Poseen el nivel de Inglés adecuado para utilizarlo de la manera requerida por sus actividades en la Organización*" aumenta de 0.013 a 0.227.

Conclusión:

Puesto que la matriz de datos con que se trabajó contiene variables aleatorias dependientes, es decir están correlacionadas, los valores

estimados por medio del método de imputación por regresión tienden al valor observado, por lo que se puede comprobar que este método es preferible al de la media. Además como la cantidad de datos faltantes es del 5% el método de eliminación por filas, no afecta mayormente a la matriz de varianzas y covarianzas de correlaciones.

Anexo 2

Algoritmo del Método de Imputación por Regresión (Matlab 6.5)

```
function metodo2=imputacion_regresion(datos,tol);
datos_orig=datos;
[n,m]=size(datos);
completos=zeros(n,m);
incompletos=zeros(n,m);
ind_c=0;
ind_i=0;
for fil=1:n
   contador=0;
   for col=1:m
      if datos(fil,col)==-99
         contador=contador+1;
      end
   end
   if contador>0
     ind_i=ind_i+1;
     incompletos(ind_i,:)=datos(fil,:);
  else
     ind_c=ind_c+1;
     completos(ind_c,:)=datos(fil,:);
   end
end
for col=1:m
   col_ind=1;
   dependiente=zeros(1,1);
   independientes=zeros(1,1);
   contador=0;
   for fil=1:ind_i
    if incompletos(fil,col)==-99
       contador=contador+1;
    end
   end
   if contador>0
      %hacer regresion
      %col es la columna dependiente
      for ii_fil=1:ind_c
        dependiente(ii_fil,1)=completos(ii_fil,col);
      end
      for ii_col=1:m
         if ii_col~=col
           col_ind=col_ind+1;
           for ii_fil=1:ind_c
              independientes(ii_fil,col_ind)=completos(ii_fil,ii_col);
              independientes(ii_fil,1)=1;
           end
         end
      end
      b=(inv((independientes')*independientes)*(independientes'))*dependiente;
      for aa_fil=1:ind_i
         vector_x=zeros(1,1);
         if incompletos(aa_fil,col)==-99
            ccc_col=0;
            for aa_col=1:m
               if incompletos(aa_fil,aa_col)~=-99
                 ccc_col=ccc_col+1;
                 vector_x(1,ccc_col)=incompletos(aa_fil,aa_col);
               end
            end
            estimado=b(1,1);
            nb=length(b);
            for e_ind=2:nb
               estimado=estimado+b(e_ind)*vector_x(e_ind-1);
```

```
            end
            incompletos(aa_fil,col)=estimado;
          end
        end
      end
end
xx_auxfil=0;
for xx_fil=1:ind_i
   xx_auxfil=xx_auxfil+1;
   datos(xx_auxfil,:)=incompletos(xx_fil,:);
end
for xx_fil=1:ind_c
   xx_auxfil=xx_auxfil+1;
   datos(xx_auxfil,:)=completos(xx_fil,:);
end

%proceso iterativo
diferencia=100000;
%while diferencia>tol

%datos

iteraciones=1;
%while iteraciones<5
datos_anterior=datos;
iteraciones
datos

while diferencia>tol
    datos_anterior=datos;
    iteraciones=iteraciones+1;
   for yy_col=1:m
      col_ind=1;
      contador=0;
      dependiente=zeros(1,1);
      independientes=zeros(1,1);
      for yy_fil=1:n
         if datos_orig(yy_fil,yy_col)==-99
            contador=contador+1;
         end
      end
      if contador>0
         %hacer regresion (otra vez)
         %dependiente
         for ii_fil=1:n
           dependiente(ii_fil,1)=datos(ii_fil,yy_col);
         end
         %independientes
          for ii_col=1:m
            if ii_col~=yy_col
              col_ind=col_ind+1;
              for ii_fil=1:n
                 independientes(ii_fil,col_ind)=datos(ii_fil,ii_col);
                 independientes(ii_fil,1)=1;
              end
            end
         end
         b=(inv((independientes')*independientes)*(independientes'))*dependiente;

         %calcular estimado
         for aa_fil=1:n
            vector_x=zeros(1,1);
            if datos_orig(aa_fil,yy_col)==-99
               ccc_col=0;
               for aa_col=1:m
                  if datos_orig(aa_fil,aa_col)~=-99
                    ccc_col=ccc_col+1;
```

```
                vector_x(1,ccc_col)=datos(aa_fil,aa_col);
              end
            end
            estimado=b(1,1);
            nb=length(b);
            for e_ind=2:nb
              estimado=estimado+b(e_ind)*vector_x(e_ind-1);
            end
            datos(aa_fil,yy_col)=estimado;
          end
        end
      end
    end
  %verificar tolerancia
  maximo=0;
  for mm_fil=1:n
    for mm_col=1:m
      if abs(datos(mm_fil,mm_col)-datos_anterior(mm_fil,mm_col))>maximo
        maximo=abs(datos(mm_fil,mm_col)-datos_anterior(mm_fil,mm_col));
      end
    end
  end
  diferencia=maximo;
  iteraciones

  datos
end
%----------------
disp('La matriz de datos converge luego de ')
iteraciones
disp('iteraciones, la matriz resultante es la siguiente:')
datos
metodo2=datos;
```

Referencias Bibliográficas

[1] Azarang, M. & García, E (1996) "***Simulación y Análisis de Modelos Estocásticos",*** Editorial McGraw-Hill Interamericana Editores, México-México.

[2] Coss, R. (1991) ***"Simulación",*** Un enfoque práctico, Editorial Limusa, México-México.

[3] Freund, J., Miller, I., Miller, M. (2000) ***"Estadística Matemática con Aplicaciones"***, Editorial Pearson Educación, México D.F., México.

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

[5] Mendenhall, W., Wackerly, D., & L-Scheaffer, R. (2002) ***"Estadística Matemática con aplicaciones",*** Thomson, Sexta Edición, México-México.

[6] Rencher, A (1998) ***Multivariate Statistical Inference and Aplications,*** Wiley Series in Probability and statistics, New York-United States of America.

[7] 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.

[8] Pérez, C. (2000) ***Técnicas de Muestreo Estadístico,*** Teoría y Práctica y Aplicaciones Informáticas, Editorial Alfaomega, Madrid-España.

[9] Gómez, J. & Palarea, J (2003) "***Inferencia basada en imputación múltiple en problemas con información incompleta",***http://www.udc.es/dep/mate/biometria2003/Archivos/ot83.pdf, Fecha de Última Visita: febrero de 2006, Guayaquil-Ecuador.

[10] Herrero, F. & Cuesta, M (2004) "***Introducción al Álgebra Matricial***",http://www.psico.uniovi.es/Dpto_Psicologia/metodos/tutor.3/vector.html, Fecha de Última Visita: marzo de 2006, Guayaquil-Ecuador.

[11] Kennedy, W & Gentle, J. "*Generación de números aleatorios*" http://math.uprm.edu/~edgar/LEC9COMP.PDF, Fecha de Última Visita: abril de 2006, Guayaquil-Ecuador.

[12] López, V. (2005) ***"Comparación de los métodos de imputación con respecto al poder de separación del modelo de regresión logística"***, http://grad.uprm.edu/tesis/lopezvazquez.pdf, Fecha de Última Visita: febrero de 2006, Guayaquil-Ecuador.

[13] Tarifa, E. (2002) ***"Teoría de Modelos y Simulación"*** http://www.modeladoeningenieria.edu.ar/unj/tms/apuntes/cp3.pdf, Fecha de Última Visita: marzo de 2006, Guayaquil-Ecuador.