CAPÍTULO III

**3. TÉCNICAS DE IMPUTACIÓN APLICABLES**

* 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 , variable por variable. Supongamos que para una variable  tenemos registrados  de los  valores investigados y () “datos faltantes”, por lo que para los () datos, los valores a ser imputados en la variable  se determinan así:



(3.1)

Siendo  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 la segunda variable la tercera variable  y la cuarta variable , esto es:

,

**Primer Caso: Falta *un dato* en solo una variable**

Se supone que la variable aleatoria *X4* que proviene de una distribución Poisson con , tiene un valor faltante, el, 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 | | | |
| *X1* | *X2* | *X3* | *X4* |
| 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 *X4*, con el dato faltante es , entonces reemplazamos en , 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:



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



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 *X4,* () 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 *X4* con un valor completado es igual a:





La media para los “datos incompletos” también es igual: 

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** | *X1* | *X2* | *X3* | *X4* | | *X1* | 2.044 |  |  |  | | *X2* | 0.067 | 2.044 |  |  | | *X3* | -0.533 | -0.356 | 8.900 |  | | *X4* | -0.267 | -0.844 | 3.033 | **11.878** | |
| **Matriz de Varianzas y Covarianzas**  **(Un Dato Completado con Imputación en** *X4***)**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Variables** | *X1* | *X2* | *X3* | *X4* | | *X1* | 2.044 |  |  |  | | *X2* | 0.067 | 2.044 |  |  | | *X3* | -0.533 | -0.356 | 8.900 |  | | *X4* | 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 *X3* y *X4* , muestran la mayor covarianza (3.033), seguida por la covarianza entre *X2*y *X4* (-0.844). También se aprecia un valor “grande” en la varianza de la variable *X4* (11.878), por ende valores de esta variable tienden a distribuirse lejos de la media, mientras que las variables *X1* y *X3*  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 *X4*, 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 *X3*y *X4* 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 , datos que provienen de una distribución Poisson con ; faltan:  y . 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 | | | |
| *X1* | *X2* | *X3* | *X4* |
| 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 *X1*, con los dos datos faltantes es, entonces reemplazamos en .

El vector de medias con dos datos completados en *X1* es:



|  |  |
| --- | --- |
| **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** | | **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 |   **Estimadores** | **Diagrama de Cajas** |

**Elaborado por**: G. Cuenca

Podemos apreciar en Cuadro 3.3 que la mediana de los datos con imputación en la variable *X1* 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** | *X1* | *X2* | *X3* | *X4* | | *X1* | **2.044** |  |  |  | | *X2* | 0.067 | 2.044 |  |  | | *X3* | -0.533 | -0.356 | 8.900 |  | | *X4* | -0.267 | -0.844 | 3.033 | 11.878 | |
| **Matriz de Varianzas y Covarianzas**  **(Dos datos completados con Imputación en** *X1***)**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Variables** | *X1* | *X2* | *X3* | *X4* | | *X1* | **1.728** |  |  |  | | *X2* | -0.211 | 2.044 |  |  | | *X3* | -0.436 | -0.356 | 8.900 |  | | *X4* | -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 *X1* con las demás variables disminuyó, por ejemplo la covarianza entre *X1*  y *X2* , 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 , datos que provienen de una distribución Poisson con ; faltan:,  y  y dos en la variable , datos que provienen de una distribución Poisson con ; faltan: y . 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 | | | |
| *X1* | *X2* | *X3* | *X4* |
| **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  y 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 | | | |
| *X1* | *X2* | *X3* | *X4* |
| **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:



|  |  |
| --- | --- |
| **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*** *X1***”**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **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*** *X1***”** |
| **Estimadores “*Variable*** *X3***”**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **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*** *X3****”*** |

**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** | *X1* | *X2* | *X3* | *X4* | | *X1* | **2.044** |  |  |  | | *X2* | 0.067 | 2.044 |  |  | | *X3* | -0.533 | -0.356 | **8.900** |  | | *X4* | -0.267 | -0.844 | 3.033 | 11.878 | |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Variables** | *X1* | *X2* | *X3* | *X4* | | *X1* | **0.762** |  |  |  | | *X2* | -0.032 | 2.044 |  |  | | *X3* | 0.294 | -0.250 | **8.542** |  | | *X4* | 0.159 | -0.844 | 1.750 | 11.878 |   **Matriz de Varianzas y Covarianzas**  **(Datos completados con Imputación en** *X1 y X3* **)** |

**Elaborado por**: G. Cuenca

La covarianza entre *X1* y *X4,* se incrementa de -0.267 en los datos originales a 0.159 en los datos completados con imputación.

El valor de la covarianza entre *X3* y *X4* 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 *p-*1 variables independientes  es un modelo del tipo:

; ~ 

(3.2)

Donde  son las variables explicativas de la regresión, es la variable explicada y  es el ruido o error aleatorio.

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



Donde  es el valor de la *j*-ésima variable independiente, para la *i*-èsima observación 

Los resultados se facilitan con el uso de notación matricial con , de la siguiente manera:

   

Por tanto, las  ecuaciones que representan las  como función de las , los estimadores  y los  se pueden escribir como:



(3.3)

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



   

Dado que

= 

(3.4)

y

=

(3.5)

Vemos que las ecuaciones de Mínimos Cuadrados están dadas por

(3.6)



En consecuencia,

(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 , es el coeficiente de correlación lineal entre *X* y *Y* , pero en el cual ya se considera a *X* como una variable aleatoria*.*



(3.8)

En el modelo de regresión lineal estudiado , *X*  es un valor fijado por el investigador, es decir no es variable aleatoria.

(3.9)



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

 donde 

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 , 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  es igual a cero. Se debe recordar que  es positivo si *X* y *Y*  tienden a aumentar juntas y es negativo si  disminuye a medida que aumenta *X.*

**3.3.3 Imputación por Regresión**

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

Supongamos que  es el único valor faltante en la entrada de la *i*-ésima fila **X ,**luego usamos los datos en la sub-matriz con las (*p*-1) filas completas,  se retrocede en las otras variables para obtener la ecuación de predicción

(3.10)



El cálculo de los coeficientes de la regresión como explicamos en la sección anterior es de la forma:



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. 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  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  y  conocidos, , *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 | | |
| *X1* | *X2* | *X3* |
| 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

**Elaborado por:** G. Cuenca

|  |  |
| --- | --- |
| **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** | *X1* | *X2* | *X3* | | *X1* | 140.509 |  |  | | *X2* | 49.759 | **72.207** |  | | *X3* | 1.948 | 3.677 | **0.250** | | **Matriz de Correlaciones**  **(Datos Originales)**   |  |  |  |  | | --- | --- | --- | --- | | **Variables** | *X1* | *X2* | *X3* | | *X1* | 1.000 |  |  | | *X2* | 0.494 | 1.000 |  | | *X3* | 0.328 | **0.865** | 1.000 | |

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 *X1* y *X2*, esto es 49.759 seguida por la covarianza entre *X2*y *X3*, 3.677.

En la matriz de correlaciones, se nota que la mayor correlación se da entre las variables *X2* y *X3* (0.865), seguida por 0.494 entre las variables *X1* y *X2* .

**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** | | |
| *X1* | *X2* | *X3* |
| 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 *X1*  y *X3* son las variables independientes que van a explicar a *X2*; para las observaciones tercera a la décima, utilizando la ecuación de regresión ;

|  |
| --- |
| **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: 

Los valores de los betas se los evalúa en las dos entradas de valores completos en la primera fila (*X1=*35.011, *X3=*2.801)





****

Similarmente hacemos la siguiente regresión pero ahora *X1*  y *X2* son las variables independientes que van a explicar a *X3,* donde , los que se evalúan en las dos entradas de valores completos en la segunda fila (*X1=*35.002, *X3=*4.901)





****

**3° Paso**

Ahora insertamos estos estimadores, 4.208 en **** y 3.099 en **,** en los valores faltantes y calculamos la ecuación de regresión basada en las diez observaciones. (Ver Tabla 3.7)

Utilizando la ecuación  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** | | |
| *X1* | *X2* | *X3* |
| 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



****

De igual forma obtenemos la ecuación para *X3* que nos da el nuevo valor de predicción



****

**4° Paso**

Nuevamente insertamos los estimadores calculados, 3.696 en **** y 2.978 en **** 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** | | |
| *X1* | *X2* | *X3* |
| 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



****



****

Aquí hay un cambio en las siguientes iteraciones. Estos valores (**,**) tienden a los verdaderos valores (*X2=*3.500y *X3=*2.702) que inicialmente la regresión estimó así (**** y ****). Además si se realizaba la imputación por la media los valores de *X2* y *X3*  serían  y . (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** | *X1* | *X2* | *X3* | | *X1* | 140.509 |  |  | | *X2* | 49.759 | 72.207 |  | | *X3* | 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** | *X1* | *X2* | *X3* | | *X1* | 140.509 |  |  | | *X2* | 50.300 | 71.677 |  | | *X3* | 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** | *X1* | *X2* | *X3* | | *X1* | 140.509 |  |  | | *X2* | 49.909 | 72.050 |  | | *X3* | 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** | *X1* | *X2* | *X3* | | *X1* | 140.509 |  |  | | *X2* | 49.927 | 72.032 |  | | *X3* | 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 *X1* y *X2*  se incrementa de 49.759 a 50.300, Así como también la covarianza entre *X1* y *X3*  de 1.948 a 2.251.

Mientras que las covarianzas con el segundo resultado de predicción empiezan a disminuir es decir, la covarianza entre *X1* y *X2*  diminuye de 50.300 a 49.909.

La covarianza entre *X1* y *X2*  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  y  conocidos, , *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 | | |
| *X1* | *X2* | *X3* |
| 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 ,  y  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 | | |
| *X1* | *X2* | *X3* |
| 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 *X2*  y *X3* son las variables independientes que van a explicar a *X1*; utilizando la ecuación de regresión 

|  |
| --- |
| **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** *X1***)** |
| **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 



De manera similar hacemos la siguiente regresión pero ahora *X1*  y *X3* son las variables independientes que van a explicar a *X2,* donde 

|  |
| --- |
| **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



Por último hacemos regresión donde, *X1*  y *X2* son las variables independientes que van a explicar a *X3*





|  |
| --- |
| **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** *X3***)** |
| **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 **,** 7.443 en **** y 2.358 en **,** (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** | | |
| *X1* | *X2* | *X3* |
| 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







Insertamos los nuevos estimadores, 29.207 en **,** 0.020 en **** y 2.281 en **,** 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** | | |
| *X1* | *X2* | *X3* |
| 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







Los nuevos estimadores son insertados, 26.713 en **,** -2.666 en **** y 2.647 en **,** 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** | | |
| *X1* | *X2* | *X3* |
| 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







Los nuevos estimadores insertados son, 21.405 en **,** -0.047 en **** y 2.687 en **,** 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** | | |
| *X1* | *X2* | *X3* |
| 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







Los nuevos estimadores que se insertan son, 22.199 en **,** 0.275 en **** y 2.749 en **,** 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** | | |
| *X1* | *X2* | *X3* |
| 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







Entonces los estimadores que se insertarán son, 22.316 en **,** 1.152 en **** y 2.772 en **,** 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** | | |
| *X1* | *X2* | *X3* |
| 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







Continuamos insertando los estimadores que ahora, 22.974 en **,** 1.583 en **** y 2.814 en **.** (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** | | |
| *X1* | *X2* | *X3* |
| 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







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 **,** 6.382 en **** y 2.347 en **.**

|  |
| --- |
| **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 (****= 35.002*,* ****= 4.901*y* ****= 2.702), y el último resultado de predicción por medio del método de imputación por regresión (****= 29.547*,* ****= 6.382 *y* ****= 2.347), es el siguiente:

****

****

****

Mientras que el error entre el dato observado de cada variable (****= 35.002*,* ****= 4.901*y* ****= 2.702), y el dato estimado por medio del método de imputación por la media (****= 27.371*,* ****= 7.445*y* ****= 3.134), es el siguiente:

****

****

****

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** | *X1* | *X2* | *X3* | | *X1* | 140.509 |  |  | | *X2* | 49.759 | 72.207 |  | | *X3* | 1.948 | 3.677 | 0.250 | |
| **Matriz de Varianzas y Covarianzas**  (Datos Completados por la Media en *X1, X2 y X3* )   |  |  |  |  | | --- | --- | --- | --- | | **Variables** | *X1* | *X2* | *X3* | | *X1* | 134.685 |  |  | | *X2* | 51.700 | 71.560 |  | | *X3* | 2.277 | 3.567 | 0.232 | |
| **Matriz de Varianzas y Covarianzas**  (Datos Completados por Regresión en *X1, X2 y X3*)   |  |  |  |  | | --- | --- | --- | --- | | **Variables** | *X1* | *X2* | *X3* | | *X1* | 135.159 |  |  | | *X2* | 51.469 | 71.673 |  | | *X3* | 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 *X1* (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 *X3*  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 *X2*y *X3* 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 *X2*y *X3* donde su valor aumentó de 3.567 a 3.672.