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.