Terra. Vol. XXII, No. 31, 2006, pp. 127-152.

UNA INTRODUCCIÓN A LA IMPUTACIÓN DE VALORES PERDIDOS

AN INTRODUCTION TO THE IMPUTATION OF LOST VALUES

Lelly Useche

Dulce Mesa

RESUMEN:

Definiciones básicas acerca del tratamiento de la no respuesta, principalmente técnicas de imputación, son explicadas en este artículo. Se destacan dos partes; la primera, una revisión relacionada con los problemas más generales de la no respuesta, los principales investigadores en esta área, sus avances, las precauciones a tomar en cada una de las fases de la investigación para disminuir los errores en ellas, así como sus causas y consecuencias. La segunda, relacionada directamente con las técnicas de imputación; su objetivo, sus ventajas y desventajas, se presentan los diferentes tipos de técnicas de imputación, se discute la manera de seleccionar la técnica adecuada, de igual forma, los pasos a seguir así como la evaluación de estas técnicas.

PALABRAS CLAVE*:* No respuesta, imputación, revisión.

****

ABSTRACT:

Basic definitions regarding the treatment of a "non-answering position" specially imputation techniques are explained in this article. Two different sections are brought to light: The first, a revision related to the general problems of "non-answering", the main researchers in this area, their main advances, their preccautions in undertaking each research phase in order to avoid error and also their causes and consequences. The second is directly related to imputation techniques; their objective, their advantages and their disadvantages. Different types of imputation techniques are presented and means of choosing adequate techniques are discussed. Equally, the steps followed to evaluate these techniques are also discussed.

KEY WORDS: revision, no answer, imputation, technique.

****

INTRODUCCIÓN

En la mayoría de los estudios muestrales y/o censales, principalmente en la medición de unidades, encontramos múltiples obstáculos, tales como, perder una medición, lo que genera espacios vacíos, que producen problemas en el análisis posterior.

Desde hace ya varias décadas, se ha venido estudiando la forma de "llenar" estos espacios vacíos, con el fin de obtener un conjunto de datos completos para analizarse por la vía de los métodos estadísticos tradicionales. Sin embargo, esta situación se complica cuando se presentan en una matriz de datos formada por diversas variables sobre la cual se realizan estudios multivariantes, haciéndose necesario la aplicación de métodos que convenientemente imputen conjuntamente los datos.

En los últimos tiempos, se han desarrollado con la ayuda del avance de la computación, nuevas formas de estudiar los datos faltantes multivariantes, obteniéndose una variedad de técnicas basadas en diferentes enfoques según las características de la data. Aún así, todavía son muchas las deficiencias que enfrentan las técnicas actuales y que son necesarias resolverlas, como los sesgos en las estimaciones, alteración de la relación entre las variables, cambios en las varianzas, entre otros.

PROBLEMAS GENERALES DE LA NO RESPUESTA

Cuando se aplica una encuesta, es muy probable que nos encontremos con espacios en blanco en el área de respuesta que denominamos "no respuesta". Su origen puede estar en un "no lo sé" o en expresiones de valores o frases incongruentes con la pregunta planteada, debido posiblemente a inconvenientes que se presentan para lograr la medición en el momento preciso, o son mediciones muy difíciles de obtener, o porque la pregunta o medida lleva consigo ciertos compromisos que prefieren mantenerse sin responder. Esta variedad de situaciones llevan a la "no respuesta", también conocida como datos faltantes.

La ausencia de datos plantea un gran problema que enfrenta el analista. En ocasiones, por ejemplo, con grandes conjuntos de datos y con poca proporción de pérdida, pudiera ignorarse la ausencia de datos, pero esto no es conveniente cuando se trata de pocas observaciones o de altas proporciones de pérdida.

La proporción de ausencia de ítems en un registro puede variar; dependiendo del estudio, de la dificultad de llenar el cuestionario o de la medición de una unidad, sin embargo, dependerá del investigador

considerar el registro como pérdida parcial o como pérdida total cuando el número de ítems perdidos sea considerable. Otra situación cuestionada por el investigador es determinar hasta qué porcentaje de pérdida de ítems se considera tratable mediante imputación o considerarlo como una mala recolección de información, y por tanto, la base de datos obtenida es muy defectuosa que simplemente no debe ser usada. La respuesta a esto no es tan sencilla. En la práctica se habla de pérdidas máximas entre 1 y 20% de la data dependiendo de la exactitud del estudio y del área de investigación entre otros factores. Por ejemplo, en las ciencias médicas, la precisión es un factor muy importante en la obtención de resultados y estudios profundos, no pueden permitir la imputación de muchos valores que nunca serán reales, sólo para poderlos analizar, mientras que en las ciencias sociales permite porcentajes de ítems imputados más altos, siempre estará en manos del investigador esta decisión.

Cuando no se pueden ignorar los datos faltantes, la manera más adecuada de tratarlos es llenar esos espacios faltantes con valores plausibles; a este procedimiento es lo que denominamos imputación.

Lohr (1999), refiriéndose a la importancia de los procedimientos de imputación señala que ésta no radica sólo en reducir el sesgo por las ausencias de respuestas, sino también para producir un conjunto de datos rectangulares y "limpios" sin datos faltantes.

Son muchas las técnicas de imputación que han surgido, sobre todo desde la década de los setenta, que emplean enfoques univariantes y multivariantes. Se han empleado enfoques basados en modelos como: funciones de verosimilitud, regresión y descomposición de matrices en valores singulares, entre otros. A pesar de estos avances, no se ha encontrado una metodología capaz de reproducir la data o que pueda resolver en forma totalmente satisfactoria el tratamiento de los datos faltantes, debido, generalmente, a problemas en cuanto a las alteraciones

de la distribución de los datos, alteración en la relación de las variables, sesgo en las estimaciones, inflación de la varianza, entre otros, razón por la cual, aún se sigue investigando en busca de mejorar las técnicas existentes.

FUENTES DE ERROR EN LAS INVESTIGACIONES ESTADÍSTICAS

Las investigaciones estadísticas pueden ser; investigaciones por muestreo, investigaciones exhaustivas, experimentos comparativos y estudios observacionales. Todas estas investigaciones son estructuradas por diversas fases en la cual pueden estar presentes fuentes de errores, por ejemplo; en la fase de planteamiento de la investigación, el no entender de manera adecuada el objetivo de la investigación, lleva a una mala definición de la población, y por tanto, al uso de un instructivo de recolección de información incorrecto, definición errónea de conceptos que traería como consecuencia que las preguntas o mediciones no sean las adecuadas con la unidad de análisis, lo que podría llevar a obtener errores en las investigaciones.

En la fase de elaboración de instrumentos básicos, se pueden presentar problemas como, preguntas del cuestionario que no reflejen los objetivos de la investigación, posiciones inadecuadas de las preguntas, mala redacción, influencia por parte del encuestador, cuestionarios muy largos, preguntas que requieren de memoria a mediano o largo plazo o que sean comprometedoras (personales), lo que llevaría a que el encuestado no responda todas las preguntas del mismo.

En la tercera fase, diseño de la encuesta, puede ocurrir la no respuesta, por un diseño no acorde con el objetivo, marcos imperfectos, selección a juicio, aún cuando se propone una selección probabilística.

En la siguiente fase, organización y ejecución de operaciones de campo, la no respuesta depende mucho del comportamiento del encuestador, de una mala selección, mal entrenamiento, sentimientos personales del entrevistador que pueden ser transmitidos a los encuestados produciéndose errores, supervisión inadecuada, cobertura incompleta de la zona geográfica en estudio por razones de costo, tiempo muy corto o por errores de selección de las unidades de muestreo en campo.

Para la fase de procesamiento de datos, puede haber ausencia o inconsistencias, debido posiblemente a un perfil o capacitación inadecuada de los codificadores / transcriptores, produciéndose errores tales como; incorrecta trascripción y codificación de los datos, utilización no adecuada de códigos y validaciones, sistemas de control de calidad deficientes y uso de "software" y "hardware" inadecuados.

En cuanto a la fase de análisis de resultados, resaltan más las consecuencias de la no respuesta, debido a que ésta puede conducir a errores en la misma, ya que no contiene todas las variables necesarias en el análisis, conduciendo a un cálculo inadecuado de pesos, lo que produce una distorsión de los resultados reales.

Finalmente en la última fase llamada plan de difusión, los resultados obtenidos pueden ser erróneamente publicados, presentando excesiva información que puede confundir al lector.

TIPOS DE ERRORES PRESENTES EN UNA INVESTIGACIÓN ESTADÍSTICA

Las fuentes de errores presentes en las diferentes fases de una investigación estadística, pudieran producir dos tipos de errores; errores muestrales y errores no muestrales, o también llamados errores ajenos al muestreo.

Los errores muestrales son los errores producidos al observar una muestra de la población y no la totalidad de ella. Este tipo de error está compuesto por la variabilidad del estimador ante muestras repetidas y su sesgo, llamado sesgo técnico. (Mesa, 2004).

Los errores no muestrales son aquellos errores presentes en una investigación, no atribuibles al observar una muestra. Pueden ser aleatorios o sistemáticos.

Otra forma de clasificar los errores no muestrales es según su fuente; los cuales pueden ser:

- Errores de cobertura: son producidos por problemas en el marco muestral; por la no inclusión de algunas unidades de observación, quedando excluidas del proceso de selección, y por ende tienen probabilidad nula de ser seleccionadas. Otro problema puede ser por exceso o duplicación, el cual ocurre cuando algunas unidades de observación aparecen más de una vez en el marco de muestreo.
- Errores de respuesta: ocurren cuando la información que se obtiene de la unidad de observación es incorrecta, estos errores se producen en la fase de recolección de datos y no en la fase de procesamiento de datos.
- Errores de falta de respuesta: surgen cuando las unidades de observación seleccionadas para la encuesta no proporcionan todos los datos que deberían recogerse.

Tipos de no respuesta

La no respuesta, aunque cuando no se quiera, siempre estará presente en toda investigación que involucre medición, de la cual, muchas veces no se obtendrán registros completos por diversas causas ajenas al investigador.

La situación idónea en una investigación es obtener una base de datos completa, de valores reales que permita llevar a cabo las tradicionales técnicas de análisis de datos.

La no respuesta, puede presentarse de dos maneras:

- La no respuesta total, es cuando falta todo el registro de una base de datos, por ausencia de la unidad a medir o por impedimento de efectuar un conjunto total de mediciones de variables en un determinado momento específico, es decir, no se recoge ningún dato de la unidad de la muestra. Por ejemplo, cuando se lleva a cabo la aplicación de encuesta en hogares y en algunas viviendas seleccionadas, y no se encuentran personas al momento de aplicar el instrumento, generándose una pérdida total de las respuestas del cuestionario que se le iba a aplicar a ese hogar o a esa persona.
- La no respuesta parcial, se presenta cuando hay ausencia de una o más variables, sin llegar a la ausencia completa de un registro, ejemplo; un individuo a encuestar se encuentra, pero no responde algunas preguntas del cuestionario o a una unidad no se le efectuaron algunas mediciones, por fallas en los equipos, sin embargo, otras mediciones si se llevaron a cabo.

La no respuesta parcial puede tener dos formas de presentarse; cuando las variables de un registro están ausentes, porque la data no está disponible, o cuando una variable produce una inconsistencia con el resto de las variables, entonces se dicen que están "pérdidas artificialmente" debido a que han sido eliminadas mediante un proceso de depuración, éstas ultimas serán tratadas como la primera forma.

Patrón de pérdida de respuesta

Uno de los puntos a considerar en la no respuesta, es el patrón de pérdida de los datos faltantes, ya que estos pueden influir en la selección del método de imputación. Los patrones de pérdida pueden ser ignorables o no ignorables.

Los patrones de pérdida pueden ser ignorables si ocurren de manera completamente aleatoria (MCAR, Missing Completely At Random) o de manera aleatoria (MAR, Missing At Random). El primer caso (MCAR), ocurre cuando la ausencia de información depende de alguna variable presente en la matriz de datos ya sea x o y. Para el segundo caso (MAR), ocurre cuando la ausencia de los datos depende de variables presentes en la matriz de datos, excluyendo la variable perdida.

Los patrones de pérdida no ignorables (NMAR) son los que ocurren cuando la ausencia de los datos depende de la variable perdida, esto traería como consecuencia estudiar el patrón de pérdida de los datos ausentes para luego imputar tomando en cuenta dicho patrón.

Causas y consecuencias de la no respuesta

Existen múltiples causas que producen constantemente la no respuesta, ya sea total o parcial, las más comunes son; las unidades de las muestras son inaccesibles, ausencia o impedimento para contactar a los respondientes, no hay cooperación, rechazo o incapacidad para responder, material perdido, el encuestador, además del diseño, la extensión, el tema a tratar, el tipo y orden de las preguntas, la redacción y el vocabulario del cuestionario. Platek (1986) plantea: "… ¿qué es

entonces lo que lleva a los encuestados a no responder? Hay tres razones: violación del derecho a la intimidad, carga de respuestas y hostilidad general del Gobierno. Al tratar con éstos problemas, es importante enfocarlos desde el lado de los encuestados, no desde el lado de las ideas preconcebidas del diseñador de encuestas".

En cuanto a las consecuencias que causa la falta de respuesta tenemos: resultados deficientes e incluso inválidos que puede llevar a una pérdida de toda la investigación, distorsión de las frecuencias marginales y/o conjuntas de las variables, sesgos en las estimaciones, disminución del tamaño de la muestra y todo lo que esto implica (aumento del error de muestreo, falta de representación en grupos o variables, estimaciones imposibles de obtener).

TRATAMIENTOS GENERALES DE LA NO RESPUESTA

Hay varias maneras de tratar la no respuesta, según Sande (1982), estas pueden ser:

- Eliminar todos los registros que tengan al menos un dato faltante, lo cual puede ser factible si tenemos un gran conjunto de datos y la pérdida es ignorable o si la proporción de registros a eliminar es muy pequeña, en caso contrario, se estaría perdiendo mucha información que puede ser importante.
- Crear una categoría como "no respuesta" en aquellas variables donde hay ausencia de información, pero esto no considera información parcial que puede servir para otros análisis o para otras variables.
- Ignorar los datos faltantes en cada caso, ponderando la variable convenientemente, pero se debe tener precaución, ya que se puede llegar a base de datos inconsistentes, es decir, valores

en diferentes variables en un mismo registro que no tengan sentido que ocurran.

- Imputar los datos faltantes.

La evolución de los estudios de la no respuesta

Los primeros aportes en imputación se realizaron en 1932 por Wilks, quien propuso el reemplazo de los datos faltantes por la media de los datos presentes de la variable. Este método, puede ser aplicado cuando existen pocos datos faltantes, ya que tienden a distorsionar la distribución de las variables.

Con el avance tecnológico de los sistemas computacionales se iniciaron investigaciones sobre el particular en las décadas de los setenta y ochenta. Entre los principales autores de los últimos años que han hecho grandes investigaciones referentes a imputación figuran Kalton, Kasprzyk, Rubin, Little y Sande. Entre sus principales aportes se encuentran los siguientes:

Rubin (1976), hizo una distinción entre MAR (valores perdidos o faltantes de manera aleatoria) y MCAR (valores perdidos o faltantes de manera completamente aleatoria). En el caso de MAR, los datos perdidos dependen de los valores observados, pero no de las variables perdidas propiamente, es decir, la pérdida de información de la variable no depende de ella misma, mientras que en MCAR los datos perdidos no dependen de otros valores observados ni de otros datos perdidos. Ambas distinciones corresponden a patrones de pérdida ignorables, existiendo además de este, un segundo enfoque el cual corresponde a patrones de pérdida no ignorables, que como se mencionó anteriormente, no ocurren de manera aleatoria, sino que siguen un patrón sistémico específico, el cual hay que estudiar previamente a la imputación.

Kalton y Kasprzyk (1982), estableció las diferencias existentes entre las técnicas de ajuste ponderado, cuando hay pérdida de un registro completo llamado no respuesta total, y las técnicas de imputación para los casos de registros con pérdidas sólo de algunas variables, es decir, no respuesta parcial, estableciendo los principales efectos de ambos métodos en las respuestas, como; aumento de sesgo, principalmente en las estimaciones de ajuste ponderado y pérdida de relación entre variables al aplicar las técnicas de imputación.

Rubin (1983), expuso otros enfoques para el estudio de los datos faltantes, los cuales los clasificó como enfoque basado en la aleatorización, frecuentista, y enfoque basado en el modelo de superpoblaciones, enfoque Bayesiano. Rubin estableció algunas ventajas y desventajas entre los diferentes enfoques ante un conjunto determinado de datos, tales como; el enfoque aleatorio es más simple de computar y es más robusta, sin embargo, el enfoque Bayesiano obtiene más precisión del intervalo de probabilidad para la variable a imputar.

Little y Rubin (1987), desarrollan una nueva técnica llamada imputación múltiple, en la que los datos faltantes, son sustituidos por $m>1$ valores simulados.

La imputación múltiple permitió hacer un uso eficiente de los datos, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no-respuesta parcial introduce en la estimación de parámetros (Goicoechea, 2002).

Helmel (1987), aportó al tratamiento de la no respuesta un método llamado Listwise, el cual, es usado cuando se tiene un gran conjunto de datos y se puede eliminar la fila o columna donde se encuentra la data perdida, para obtener una base, aunque más pequeña, completa. Esta técnica es comúnmente usada hoy en día, pero no es recomendable para casos de pequeños conjuntos de datos, por la pérdida de información que ocasiona.

En la década de los 90, Todeschini (1990) propuso un k-vecino más cercano como método de estimación de valores perdidos; obteniéndose buenos resultados cuando se cuenta con información auxiliar.

Otras investigaciones han buscado maneras de mejorar las técnicas de imputación o crear nuevas, como las basadas en redes neuronales (Koikkalainen, 2002), análisis factorial (Genz y Li, 2003), en análisis de componentes principales (Gleason, y Staelin, 1975), o basada en la descomposición GH-Biplot (Vásquez, 1995), entre otras.

A partir del año 2000 se ha implementado el uso de árboles de clasificación como mejora de los procedimientos de imputación. Mesa, Tsai y Chambers (2000) realizaron un estudio de imputación del Censo del Reino Unido mediante el uso de árboles de clasificación, en el cual se llegó a las siguientes conclusiones; 1. Los árboles grandes no son garantía de obtener mejores imputaciones. 2. El software a usar tiene comparativamente poca importancia. 3. Los métodos de probabilidad más alta y la selección de una categoría aleatoria no mantienen la distribución de la variable, y 4. Ninguno de los métodos investigados fueron satisfactorios desde el punto de vista de recuperar los datos perdidos actuales individuales.

Aún son muchos los estudios que se deben llevar a cabo para resolver los problemas mencionados anteriormente, los cuales se presentan en la mayoría de los métodos de imputación que existen hoy en día, haciendo distinción en los diferentes enfoques que puedan existir y dependiendo además del conjunto de datos a analizar.

El objetivo de la imputación

El objetivo de la imputación es obtener un archivo de datos completos y consistentes para que puedan ser analizados mediante técnicas estadísticas tradicionales.

VENTAJAS Y DESVENTAJAS DE IMPUTAR

Cuando imputamos, logramos obtener una base de datos completa, la cual permitirá llevar a cabo metodologías de análisis de datos comunes y el uso de software tradicionales para su manejo. Si una imputación se lleva a cabo de manera adecuada, podría disminuir el sesgo, en caso de existir.

Por otra parte, el investigador debe estar consciente que el uso de imputación también puede llevar a afectar las distribuciones conjuntas, o incluso, distribuciones marginales de las variables, aunque el problema es menor si la distribución de los casos ausentes es la misma que la de los casos completos (patrón de pérdida ignorable), como se mencionó anteriormente. Si la técnica no es la adecuada, posiblemente, aumenta el sesgo, subestima o sobrestima la varianza, se obtienen datos imputados inconsistentes produciendo una base de datos no confiables, llevando a la interpretación errónea de los resultados por parte de los usuarios.

DIFERENTES TIPOS DE TÉCNICAS DE IMPUTACIÓN

Varios estudios (Goicoechea, 2002; Platek 1986; y Government Statistical Service 1996), indican que las técnicas de imputación se pueden clasificar de la siguiente manera:

1. Técnicas fundamentadas en información externa: cuando son basadas en variables relacionadas con una encuesta perteneciente a otras bases de datos o reglas previas. Entre estas se encuentran:

   a) Métodos deductivos: cuando los datos faltantes se deducen con cierto grado de certidumbre de otros registros completos del mismo caso, siguiendo algunas reglas específicas.

b) Tablas Look-up: cuando se hace uso de una tabla con información relacionada, como fuente de data externa para imputar los datos faltantes.

2. Técnicas determinísticas: cuando al repetir la imputación en varias unidades bajo las mismas condiciones, producirá las mismas respuestas.

a) Imputación de la media o modo: se llena el vacío del dato faltante de cada variable con la media de los registros no faltantes en caso de variables cuantitativas, o con la moda en caso de variables cualitativas. Tiene como desventaja la modificación de la distribución de la variable haciéndose más estrecha ya que reduce su varianza, además, no conserva la relación entre variables y se debe asumir una MAR. Su ventaja es la facilidad de la aplicación del método.

b) Imputación de media de clases: las respuestas de cada variable son agrupadas en clases disjuntas con diferentes medias, y a cada registro faltante se le imputará con la media respectiva de su grupo. Tiene las mismas desventajas que el caso anterior, pero en menor proporción por estar agrupadas. Igualmente es de fácil aplicación.

c) Imputación por regresión: se ajusta un modelo lineal que describa a y, variable a imputar, para un conjunto X de variables auxiliares que se deben disponer. Resuelve el problema de la distorsión de la distribución de la variable a imputar, pero puede crear inconsistencias dentro de la base de datos, pues podría obtenerse valores "imposibles", ya que el valor y es obtenido de variables auxiliares.

d) Emparejamiento media: se lleva a cabo el método (e) donde el valor de y (estimado) es comparado con casos completos, y el caso más cercano correspondiente provee el valor imputado y.

e) Imputación por el vecino más cercano: se identifica la distancia entre la variable a imputar y, y cada una de las unidades restantes (x o variables auxiliares) mediante alguna medida de distancia, entonces se determina la unidad más cercana a y, usando el valor de esta unidad cercana para imputar el faltante.

f) Algoritmo EM (Expectation Maximization): basada en la función de máxima verosimilitud, permite obtener estimaciones máximo verosímiles (MV) de los parámetros cuando hay datos incompletos con unas estructuras determinadas. Resuelve de forma iterativa el cálculo del estimador máximo verosímil mediante dos pasos en cada iteración (Little y Rubin, 1987). Este algoritmo tiene la ventaja de que puede resolver un amplio rango de problemas, incluyendo problemas no usuales que surgen de la pérdida o data incompleta, como lo es la estimación de los componentes de la varianza.

g) Redes Neuronales: son sistemas de información procesados, que reconocen patrones de los datos sin algún valor perdido para aplicarlo a la data a imputar. Estas redes son más usadas para variables cualitativas que cuantitativas, siendo más adecuadas cuando la distribución es no lineal. No es aconsejable cuando hay registros atípicos que distorsionan la red. Son costosos y requieren de capacitación del analista así como de "software" adecuado.

h) Modelos de series de tiempo: se asume que la data perdida ocurre de tal forma, y en tal sistema, que el problema se reduce a una situación, en la cual, hay una serie de tiempo, donde una(s) serie(s) de observaciones están perdidas, haciendo óptimo el uso de interrelaciones entre sucesivas

observaciones en cada serie de tiempo, mediante el uso de un modelo adecuado para estas series.

3. Técnicas aleatorias o estocásticas: son aquellas que cuando se repite el método de imputación bajo las mismas condiciones para una unidad, producen resultados diferentes.

   a) Imputación aleatoria de un caso seleccionado: para cada caso con una celda faltante, se selecciona un donante aleatoriamente para ser asignado al dato faltante.

   b) Imputación aleatoria de un caso seleccionado entre clases: se realiza de igual forma que para el caso (k) pero se lleva a cabo dentro de clases previamente creadas.

   c) Imputación secuencial Hot-Deck: cada caso es procesado secuencialmente. Si el primer registro tiene un dato faltante, este es reemplazado por un valor inicial para imputar, pudiendo ser obtenido de información externa. Si el valor no está perdido, éste será al valor inicial y es usado para imputar el subsiguiente dato faltante. Entre las desventajas se encuentra que cuando el primer registro está perdido, se necesita de un valor inicial, (generalmente obtenido de manera aleatoria), además cuando se necesitan imputar muchos registros se tiende a emplear el mismo registro donante, llevando esto a su vez la pérdida de precisión en las estimaciones.

   d) Imputación jerárquica Hot-Deck: similar al método secuencial anterior. En esta se organizan dentro de clases haciendo uso de variables auxiliares en forma de una estructura jerárquica. Si el donante no es encontrado en un nivel de clasificación, las clases pueden ser colapsadas en grupos más anchos hasta que el donante sea encontrado.

e) Imputación por regresión aleatoria: se hace primero un procedimiento de regresión (e), luego un término residual es adicionado para imputar los valores de y. Este término de error puede ser obtenido de diferentes maneras, una de ellas es a través de los residuos del modelo de regresión, generado con registros completos, eligiendo uno de éstos residuos aleatoriamente.

f) Imputación por regresión logística: similar al la técnica anterior, pero para imputar variables binarias.

CÓMO SELECCIONAR LA TECNICA ADECUADA DE IMPUTACIÓN

Seleccionar un método de imputación adecuado es una decisión de gran importancia, ya que para un conjunto de datos determinado, algunas técnicas de imputación podrían dar mejores aproximaciones a los valores verdaderos que otras. Para la selección de la técnica de imputación adecuada, no hay reglas específicas, dependerá entonces del tipo del conjunto de datos, tamaños del archivo, tipo de no respuesta, patrón de pérdida de respuesta, de los objetivos de la investigación, características específicas de la población, características generales de la organización del estudio, software disponible (Entilge 1996), importancia de los valores agregados o de los valores puntuales (microdato), distribuciones de frecuencias de cada variable, marginal o conjunta, etc. Hay que tomar en cuenta que muchas veces la técnica de imputación seleccionada puede ser adecuada para algunas variables pero para otras no y será decisión del investigador seleccionar la técnica que menos afecte las estimaciones de las variables.

Fellegi y Holt (1971), plantean que: "La técnica de imputación seleccionada debe superar las reglas de validación, cambiando lo menos

posible los registros, manteniendo la frecuencia de la estructura de la data".

Goicoechea (2002), resume los criterios a tomar en consideración para seleccionar la técnica adecuada para imputar.

1. Tipo de variable a imputar: si es continua, tomar en cuenta el intervalo para la cual se define, y si es cualitativa, tanto nominal como ordinal, las categorías de las variables.

2. Parámetros que se desean estimar: si deseamos conocer sólo valores agregados como la media y el total, se pueden aplicar métodos sencillos como imputación con la media o moda, sin embargo, puede haber subestimación de la varianza. En caso de que se requiera mantener la distribución de frecuencia de la variable y las asociaciones entre las distintas variables, se deben emplear métodos más elaborados aplicando imputación de todas las variables faltantes del registro.

3. Tasas de no respuesta y exactitud necesaria: cuando el porcentaje de no respuesta es alto en una base de datos, se considera que no hay confiabilidad en los resultados que se obtengan con el análisis de esta base.

4. Información auxiliar disponible: es bueno hacer uso de la información auxiliar disponible, ya que con ella podemos deducir información de los valores ausentes de una variable o hallar grupos homogéneos respecto a una variable auxiliar que se encuentre altamente correlacionada con la variable a imputar, y de esta manera encontrar un donante adecuado que sea similar al registro receptor.

PASOS PARA LLEVAR A CABO UN PROCESO DE IMPUTACION

Según Goicoechea (2002) los pasos que se llevan a cabo para realizar imputación son:

Paso 1: una vez que se dispone de un archivo con datos faltantes, se recopila y valida toda la información auxiliar disponible que pueda ser de ayuda para la imputación.

Paso 2: se estudia el patrón de pérdida de respuesta. Posteriormente se observa si hay un gran número de registros que simultáneamente tienen no respuesta en un conjunto de variables.

Paso 3: se seleccionan varios métodos de imputación posibles y se contrastan los resultados.

Paso 4: se calculan las varianzas para los distintos métodos de imputación seleccionados con el objetivo de obtener estimaciones con el mínimo sesgo y la mejor precisión.

Paso 5: se concluye a partir de los resultados obtenidos

IMPORTANCIA DE LA CLASIFICACIÓN EN LA IMPUTACIÓN

La eficiencia del uso de una técnica de imputación, en una base de datos, muchas veces se ve afectada por la heterogeneidad de los datos, que hace que el valor donante pertenezca a un registro de una característica muy distinta a la del registro a imputar, obteniendo datos inconsistentes y/o sesgos muy grandes. Por ejemplo, si imputamos usando un valor aleatorio, se obtendrían mejores resultados si se clasificaran los datos formando grupos homogéneos y luego se seleccionara el donante dentro del grupo, el cual tendría características más similares en relación al receptor.

Formar clases de imputación podría generar buenas estimaciones de la data faltante, ya que se imputarían valores similares a los verdaderos y podría haber consistencia con los otros valores del registro. Little y Rubin (1987), exponen que existen dos requerimientos importantes para el logro de esta meta: "uno es que entre cada clase de imputación, el valor del donante representa los valores de los faltantes, ejemplo: suponer pérdida aleatoria, y la otra es que entre cada clase, el valor donante tenga una varianza pequeña". Lo ideal sería que no existiese varianza y por tanto no habría error en las estimaciones pero, ¿Cómo clasificar para satisfacer estas características?, ¿Cómo lograr donantes potenciales homogéneos respecto a la variable que se necesita imputar? Para ello existe un conjunto de técnicas de clasificación; entre ellas los árboles de decisión, en especial los árboles de clasificación y regresión (cuyas siglas en inglés son CART).

Evaluación de las técnicas

Existen muchas maneras de llevar a cabo la evaluación y validación de una o más técnicas de imputación que hayan sido aplicadas a una situación particular, la más común de ellas es el uso de simulación, que consiste en aplicar perdidas artificiales de datos a una base original, a esta base de datos con perdidas artificiales es llamada base de datos disponible. Posteriormente a esta base de datos se le suprimirán todos los registros que tengan al menos un dato faltante, obteniendo finalmente una base de datos más pequeña, pero completa.

Asumiendo que las relaciones y características de los datos perdidos es igual a la de los datos presentes (casos ignorables), se procede a aplicarle a la base de datos completa el mismo patrón de pérdida aplicado a la base de datos original.

Ante tales pérdidas, se aplica la técnica de imputación, y ésta se evalúa midiendo en ambas bases (completa e imputada) la desviación estándar, el sesgo de las estimaciones, las distribuciones de frecuencia conjuntas y marginales, entre otras magnitudes de los efectos de imputación, mediante el uso de tablas de contingencia, de distribuciones de frecuencia o prueba de hipótesis. Para ello se hará uso de estadísticos de conservación de la distribución, como el estadístico T para variables continuas y el estadístico Chi-cuadrado para variables categóricas.

En la evaluación de las diferentes técnicas también se considerará el porcentaje de pérdida de los datos, para saber hasta qué porcentaje la base de datos es confiable, observando ciertos criterios, tales como, cuántas veces ha sido usado un donante, cuantos intentos fueron requeridos para completar un registro, entre otros.

Goicoechea (2002) propone una serie de medidas deseables, al evaluar las técnicas de imputación, estas medidas son:

1. Precisión en la predicción: el valor imputado debe ser lo más cercano al valor verdadero. Para ello se hace uso de tablas de contingencia.
2. Precisión en la distribución: mantener en lo posible las distribuciones marginales y conjuntas.
3. Precisión en la estimación: producir parámetros insesgados e inferencias eficientes de la distribución de los valores reales.
4. Imputación plausible: valores aceptables al aplicarles el proceso de edición.

Consecuencias del uso de imputación

La imputación, como se ha mencionado anteriormente, no garantiza que los resultados obtenidos luego de imputar sean menos

sesgados, incluso podrían obtenerse grandes sesgos. Por otra parte, si no se evalúa correctamente la técnica de imputación aplicada, los usuarios podrían tratar con datos no reales y obtener información incorrecta.

Los recursos requeridos para llevar a cabo los procesos de imputación pueden ser altos, especialmente cuando se trabaja con grandes masas de datos y un número considerable de variables, es por ello que hay investigadores que seleccionan las variables más importantes del estudio para imputar.

Discusión final

Obtener una base de datos real completa es prácticamente imposible en la realidad, por las múltiples causas que pueden acontecer que llevan a la ausencia de datos. Para resolver el problema y poder hacer usos de las técnicas estadísticas tradicionales es necesario conocer la forma de evitar la no respuesta, el tratamiento de la no respuesta, y la manera de imputar lo cual es un arte del investigador el buen manejo de la data perdida para obtener una base de datos consistente y adecuada. El investigador puede basarse en el tipo de información que posea, el enfoque que más le convenga y el uso de simulación para evaluar las diferentes opciones de técnicas de imputación que puede usar. Lo que hay que acotar, es que siempre se debe evitar la no respuesta en la medida posible, para usar imputación sólo cuando sea necesario, pues nunca una data imputada será mejor que una data real.

Referencias bibliográficas

ELTINGE, J. (1996). *Discussion of Imputation Papers*. Recuperado 05, noviembre del 2004 en: http://www.amstat.org/sections/srms/Proceedings/papers/1996_049.pdf.

FELLEGI, I. y HOLT, D. (1976). *A systematic approach to automatic edit and imputation*. Journal of the American Statistical Association. Vol 71, 353. 17-35.

GENG, Z. y LI, K. (2003). *Factorization of posteriors and partial imputation algorithm for graphical models with missing data*. Statistics and probability letters. 64, 369-379.

GLEASON, T. y STAELIN, R. (1975). *A proposal for handling missing data*. Psychometrika. Vol 40, 2. 229-252.

GOICOECHEA, P. (2002). *Imputación basada en árboles de clasificación*. EUSTAT.

GOVERNMENT STATISTICAL SERVICE. (1996). *Report of the Task Force on Imputation*. GSS Methodology Serie No. 3. Reino Unido.

HEMEL, J. y otros (1987). *Stepwise deletion: a technique for missing data handling in multivariate analysis*. Analytical Chemical Acta 193 255-268.

KALTON, G y KASPRZYK, D. (1982). *Imputing for Missing Survey Responses*, American Statistical Association.Proceeding of the Section on Survey Reseach Methods.

KOIKKALAINEN, P. (2002). *Neural Network for editing and imputation*. University of Jyvâskylâ Finland. Recuperado 05, noviembre del 2004 en: http://erin.mit.jyu.fi/dataclean/abstracs/node33.html#SECTION00093000000000000000.

LITTLE, R. y RUBIN, D. (1987). *Statistical Analysis with Missing Data. Series in Probability and Mathematical Statistics*. John Wiley & Sons, Inc. New York.

LOHR, Sh. (1999). *Muestro: Diseño y Análisis*. Editorial Thomson.

MESA, D. (2004). *Imputación y Árboles de Decisión. Caracas, Venezuela*. Guía práctica. Postgrado en Estadística, Universidad Central de Venezuela.

MESA, D. TSAI, P. y CHAMBERS, R. (2000). *Using Tree-Based Models For Missing Data Imputation: An Evaluation Using Uk Census Data*, Reporte Técnico. Proyecto AUTIMP. Recuperado 20, septiembre del 2004 en: http://www.cbs.nl/en/service/autimp/CART-Dutch%20Data-(AUTIMP).pdf.

PLATEK, R. (1986). *Metodología y Tratamiento de la no respuesta*. Seminario Internacional de Estadística en EUSKADI. Cuaderno 10.

RUBIN, D. (1976). *Inference and missing data*. Biometrika 63. 581-592.

RUBIN, D. (1983). *Panel of incomplete data in sample surveys*. En Madow, W.G., Olkin, I. y Rubin, D.B. Incomplete Data in Simple Surveys. Vol 2,12, 123-145. Report and Case Studies. New York. Academic Press.

SANDE, I. (1982). *Imputation in Surveys: Coping with reality*. The American Statistician, Vol.36, 3:145-152.

TODESCHINI, R. (1990). *Weighted k-nearest neighbour method for the calculation of missing values*. Chenometrics and Intelligent Laboratory Systems 9.201-205.

VÁSQUEZ, M. (1995). *Aportación al Análisis Biplot: Un enfoque Algebraico*. Tesis doctoral. Universidad de Salamanca. España.

WILKS, S. (1932). *Moments and distributions of estimates of population parameters from fragmentary simple*, Annals of Mathematical Statistics, B, 163-195.

****

Lelly María Useche Castro, Ingeniero Industrial, Universidad Nacional Experimental del Táchira, (UNET). Con estudios realizados en la maestría en Gerencia de Empresa Agrícola (2002) UNET. Candidato al doctorado integrado en Estadística. FACES-UCV, (2005). Profesora de las cátedras estadística I y estadística II en la Universidad Nacional Experimental Sur del Lago, desde el año 2001. lellyuseche@cantv.net

****

Dulce Maria Mesa Ávila, Ph. D. (Social Statistics, 2002) University of Southampton, U. K. Licenciado en Ciencias Estadísticas (1992). UCV. Asesor / Consultor (Permanente). Instituto Venezolano de Análisis de Datos (IVAD). Caracas. Jefe del Departamento Diseño Estadístico y Modelos de Investigación de Operaciones. Escuela de Estadística y Ciencias Actuariales. UCV. Profesor seminario III: diseño de tesis doctorales. Programa integrado de Postgrado en Ciencias Sociales. Escuela de Estadística y Ciencias Actuariales. UCV. Ha presentado conferencias y participado en otras en el Reino Unido, Alemania y España. dmesa@ivad.com.ve