Casos con Datos Faltantes: ¿Qué Hacer con Ellos?

Rogelio Puente Díaz
Marketing Group

Personas dedicadas a la investigación enfrentan decisiones difíciles cuando analizan datos. Una de estas decisiones difíciles consiste en como manejar bases de datos con valores faltantes *(missing values)*. El propósito de este artículo es explicar de manera sencilla y práctica, presentando un ejemplo aplicado, una de las técnicas más eficaces para manejar valores faltantes llamado imputación múltiple *(multiple imputation)*. Lo que sigue a continuación es un breve análisis de las ventajas y desventajas de 3 técnicas de manejo de datos faltantes.

Técnicas para Manejar Datos Faltantes

La mayoría de los paquetes de estadística ocupan *listwise deletion* para manejar datos faltantes. *Listwise deletion* consiste en no usar los casos que tengan algún dato faltante en cualquiera de las variables involucradas en el análisis. Las ventajas de esta técnica es que arroja coeficientes consistentes y sin sesgo además de errores estándar precisos, pero el no utilizar casos que tengan datos faltantes resulta en una pérdida de información y poder estadístico (Allison, 2002). Otra opción es reemplazar los valores faltantes con el promedio. Aunque esta técnica tenga un cierto atractivo intuitivo, produce coeficientes inconsistentes y sesgados por lo que su uso no es recomendable (Allison, 2002).

Imputación múltiple produce coeficientes consistentes y sin sesgo además de permitir el uso de todos o la mayoría de los casos con valores faltantes en la base de datos. A continuación ofrezco una breve explicación de la técnica de imputación múltiple.

Imputación Múltiple

Antes de comenzar con el ejemplo aplicado es importante definir algunos conceptos. La técnica de imputación múltiple asume que los datos faltantes se originaron de manera aleatoria. Imputación múltiple consiste en imputar los valores faltantes con base en la predicción de la distribución multivariada de las variables (Schafer, 1997). Es decir, cada valor faltante es reemplazado por **m >1** posibles valores (Schafer & Olsen, 1998). Estos valores son imputados a través del uso de un algoritmo que encuentra los valores que son más posibles dada la distribución multivariada de las variables. Al imputar los valores faltantes se crea una nueva base de datos. Crear una nueva base de datos no es suficiente. Se deben de crear **m >1** bases de datos que sean relativamente independientes una de la otra. La razón de crear más de una base de datos consiste en entender que los datos faltantes son reemplazados por uno de los posibles valores. Por lo tanto, obtener y conducir el análisis únicamente en una base de datos ignora el hecho de que puede haber variación en las predicciones de los valores faltantes (esta variación se toma en cuenta en la corrección de los errores estándar). Sin embargo, es importante mencionar que crear más de 5 bases de datos no aumenta considerablemente la eficacia del proceso (Schafer, 1997). Una vez que se han creado **m > 1** bases de datos, se pueden ocupar cualquiera de los paquetes de estadística, como SPSS, SAS, o LISREL para llevar a cabo el análisis.

Los coeficientes, generados por cualquier método estadístico usando las nuevas bases de datos, poseen importantes propiedades como el ser consistentes y no sesgados, pero los errores estándar tienden a estar subestimados porque no toman en cuenta el proceso de imputación múltiple (Allison, 2002). Por lo tanto, es importante ajustar los errores estándar para que tomen en cuenta el proceso de imputaciones múltiples. El ajuste puede ser realizado a través de diferentes paquetes de estadística como NORM o manualmente ejecutando los siguientes 3 pasos:

1. Elevar los errores estándar al cuadrado y obtener el promedio.
2. Calcular la varianza de los coeficientes de interés (coeficientes beta, correlaciones, etc.).

3. Sumar los resultados del paso 1 y 2 (aplicando un pequeño factor de corrección al paso 2) y tomando la raíz cuadrada.

Ejemplo Aplicado

El ejemplo que les voy a mostrar consistió en probar un modelo motivacional con tenistas mexicanos [1]. El modelo tenía 3 variables exógenas (miedo al fracaso, perfeccionismo y esperanza) las cuales afectaban directamente 4 orientaciones de meta, las cuales a su vez influenciaban importantes variables como el esfuerzo, desempeño deportivo y la elección de retos. Para poner a prueba este modelo, usé ecuaciones estructurales tratando las variables como observadas con el paquete de estadística LISREL. La base de datos tenía 200 participantes. El porcentaje de datos faltantes era relativamente bajo, 2.35% del total de los datos, pero había 83 casos con datos faltantes.

El primer paso consistió en imputar o predecir los datos faltantes con base en los valores de las otras variables. Generé 5 bases de datos con el paquete de estadística LISREL. Después de generar las 5 bases de datos, realicé el análisis de ecuaciones estructurales 5 veces, una vez con cada base de datos. Al realizar el análisis 5 veces, una vez con cada base, uno obtiene 5 conjuntos de coeficientes. Estos coeficientes y sus respectivos errores estándar fueron analizados en el paquete de estadística NORM el cual corrige los errores estándar y toma en cuenta la variación de los coeficientes obtenidos con las 5 diferentes bases de datos. El paquete de estadística NORM me permitió determinar si los coeficientes beta y gama eran significativos una vez que la incertidumbre provocada por tener valores faltantes fue tomada en cuenta.

Imputación múltiple me permitió ocupar la información de los 83 casos que tenían datos faltantes lo cual no hubiera sido posible con el uso de otras técnicas para manejar casos con valores faltantes como *listwise deletion*. Imputación múltiple me ayudó a obtener coeficientes consistentes y sin sesgo. Aunque imputación múltiple es una técnica eficiente, también tiene algunas desventajas. Entre sus desventajas se encuentran la dificultad de implementación y que es imposible poner a prueba si el mecanismo de datos faltantes de nuestros casos es aleatorio.

En resumen imputación múltiple ofrece varias ventajas sobre las otras técnicas de manejo de datos faltantes:

1) Imputación múltiple basada en la predicción de la distribución multivariada de las variables arroja coeficientes consistentes y sin sesgo, algo que no puede hacer el reemplazo de los datos faltantes con el promedio.

2) Imputación múltiple permite la utilización de todos o la mayoría de los casos, algo que con *listwise deletion* no se puede hacer.

3) Una vez que se han generado las bases de datos, cualquier paquete de estadística como SAS y SPSS puede ser usado para realizar cualquier análisis estadístico.

El propósito de este artículo fue presentar una breve introducción del método para manejar valores faltantes llamado imputación múltiple *(multiple imputation)*. Para mayor información sobre este método, los invito a consultar las referencias.

Referencias

Allison, P. D. (2002). Missing Data. Sage: Thousand Oaks.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behavioral Research, 33, 545-571.

Notas

[1] La base de datos fue generada a través de cuestionarios autoadministrados con participantes mexicanos y fue parte de mi tesis de doctorado en la Universidad de Iowa.