Departamento Administrativo Nacional de Estadística

Dirección de Metodología y Producción Estadística – DIMPE

Metodología de Imputación
Encuesta Anual de Servicios - EAS

Julio 2004

CONTENIDO

1. MARCO CONCEPTUAL

Una gran cantidad de información, acerca de las características económicas tanto de individuos como de empresas o países, es recopilada con fines de análisis, para planear y tomar decisiones.

Al registro sistemático de mediciones u observaciones numéricas, efectuado a intervalos fijos de tiempo, se conoce como serie de tiempo y como la serie de tiempo se compone de datos numéricos, es común usar la estadística para describirla y analizarla, sea de forma descriptiva o de manera inferencial, cuyo procedimiento de esta última es la utilización de muestras, que representen a la población de estudio, para producir conclusiones válidas para toda la población.

Uno de los problemas que se presentan en el análisis estadístico inferencial es la falta de algunos registros en la serie, lo que conlleva a aumentar el error en la varianza de las estimaciones de los parámetros poblacionales; para hacer menos grave el error, se presentan dos métodos de estimación con datos faltantes que son la reponderación y la imputación.

La imputación es un método muy usado, en el cual se debe hacer el esfuerzo por imputar solo del 1% al 2% de los datos, si el porcentaje de datos imputados es muy alto se crea un error sistemático o sesgo en la varianza del estimador puntual. Pero aún si un método de imputación no produce un apreciable error, no se debe ignorar el efecto que la imputación tiene en la precisión de la varianza del estimador puntual.

La imputación es útil porque hace más viable el análisis de un conjunto de datos, asegurando consistencia en los resultados estadísticos obtenidos y reduciendo el sesgo de no respuesta.

Varias técnicas estadísticas requieren de conjuntos de datos rectangulares o en forma de matriz y que con presencia de datos faltantes, los registros se pueden restringir a un conjunto de datos completos. Esta restricción sacrifica información parcial en aquellas encuestas que no han sido diligenciadas totalmente y que se pueden utilizar o aprovechar si se hace imputación.

En la literatura estadística hay una variedad de métodos que se han propuesto para imputar datos, estos métodos son clasificados según si se genera una sola imputación para cada valor faltante (imputación simple) o se generan, bajo simulaciones, *m* imputaciones para cada valor faltante el cual genera *m* conjuntos de datos completos (imputación múltiple).

Algunos métodos de imputación usan un modelo explícito como el de una regresión ajustada, una razón o la imputación por la media. En otros métodos el modelo es implícito como el de la imputación en paquete caliente (*hot deck*) y la imputación por donadores vecinos.

2. METODOLOGÍA GENERAL DE IMPUTACIÓN DE DATOS FALTANTES

En esta metodología se utiliza la información de la muestra anual de servicios, de tal manera que los datos imputados se aproximen a los valores reales. La metodología supone que los datos de la muestra poseen autocorrelación temporal y homogeneidad en las diferentes etapas de agregación; esto significa que la imputación debe estar de acuerdo al comportamiento de la serie histórica y de los niveles que contienen al dato faltante.

Para la imputación de registros en estado de deuda, se utiliza la razón de crecimiento de los datos, en la serie, o variación de los datos presentada en la metodología de imputación de Andrés Lozano titulada "Estimación de novedades en estado de deuda", definida como:

$$\text{Variación} = \frac{X_t}{X_{t-1}}$$

Donde

X_t = dato en el período t

X_{t-1} = dato en el período anterior $t-1$

Bajo estas consideraciones, se estimará primero la variación que tendrá el dato faltante con respecto al dato del período anterior, teniendo en cuenta el comportamiento histórico de la serie de variaciones en cada empresa y el comportamiento histórico de las variaciones dentro de cada actividad, a partir de esta estimación se generará el dato faltante.

El cuadro 1 presenta un ejemplo, realizado para el establecimiento con número de orden 80020 y que pertenece a la actividad con código 5551,de las variaciones calculadas para la variable salario integral.

Cuadro1
Variaciones de la variable salario integral de 1996 a 2000

Periodo	Salario integral	Variación del salario integral	Total del salario integral por actividad	Variación Total del salario integral por actividad
1996	53048		2594984	
1997	62832	1,184437	3694641	1,423763
1998	63604	1,012287	4685218	1,268112
1999	68710	1,080278	5702401	1,217105
2000	88490	1,287877	5649922	0,990797

La variación del dato que se va a imputar se obtiene en términos de la variación histórica promedio en la empresa y en la actividad.

El modelo para imputar la variación es:

$$Var_t = \beta_1 Vac + \beta_2 Vem$$

Donde,

Var_t = Variación que se imputa en el período *t*
Vac = Variación promedio histórica dentro de la actividad
Vem = Variación promedio histórica por establecimiento.

β_i= para i = 1,2 son coeficientes de ponderación.
Cuya suma debe ser igual a uno para que haya convergencia en la imputación.

El modelo describe la imputación de la variación del dato faltante, como un promedio ponderado de las variaciones de los datos en el establecimiento y en la actividad, donde los β_i son los coeficientes de ponderación de las variaciones.

Como se expone en Lozano "el propósito es estimar los parámetros desconocidos β_i, utilizando un método iterativo con el modelo de mínimos cuadrados y restringiéndolos a que la suma sea igual a uno para que haya convergencia en la imputación."

2.1. Utilización del Modelo.

Se tendrá en cuenta los supuestos expuestos en la metodología de imputación "Estimación de novedades en estado de deuda" de Andrés Lozano (2000) para el buen desempeño del modelo, los cuales son el de homogeneidad de los datos dentro de la actividad y la autocorrelación temporal entre las variaciones de los datos de la serie histórica dentro de cada establecimiento y dentro de cada actividad, supuestos que se utilizan en la estructura del modelo.

3. EJEMPLO

Para explicar la metodología, se procederá a presentar un ejemplo de imputación de la variación y del total de la variable salario integral en el año 2000 para el establecimiento con número de orden 80020.

Utilizando la base de datos en la cual se encuentran los datos históricos por establecimiento, se calcularon las variaciones que ha tenido la variable por establecimiento y por actividad desde 1996, luego se calculó el promedio de las variaciones dentro del establecimiento y dentro de la actividad.

Empleando diferentes combinaciones de parámetros, con la restricción exigida que la suma sea igual a uno se procedió a imputar las variaciones y el total de la variable mencionada anteriormente.

En el cuadro-2 se relacionan los datos reales del total y de las variaciones de la variable dentro del establecimiento y de la actividad, datos que se quieren imputar para observar la bondad del ajuste del modelo.

Cuadro 2.
Datos de los totales y de las variaciones reales para las variables total de empleados permanentes y Producción

Salario integral	*Variación del salario integral*	*Total del salario integral por actividad*	*Variación Total del salario integral por actividad*
88490	1,287877	5649922	0,990797

Las imputaciones obtenidas para la variación y para el total de la variable salario integral se presentan en los cuadros 3 y 4 respectivamente y se muestra en la gráfica-1 el dato real y la imputación más grande y la más pequeña del salario integral.

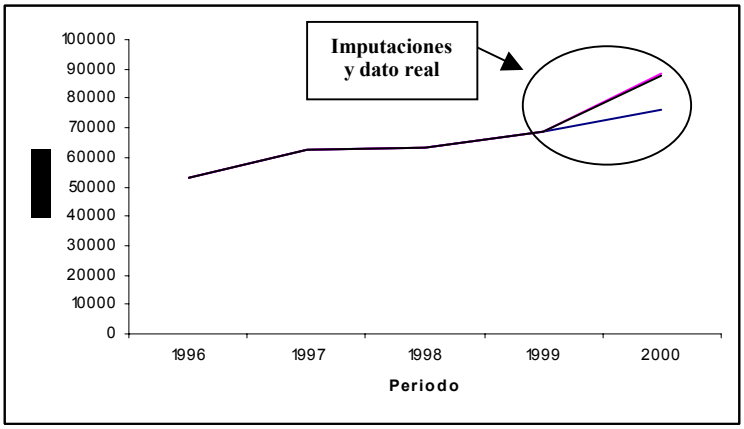
Cuadro 3.
Imputaciones de la variación del salario integral para distintas combinaciones de parámetros

	Combinación de parámetros								
	$\beta_1 = 0.1$ $\beta_2=0.9$	$\beta_1 = 0.2$ $\beta_2=0.8$	$\beta_1 = 0.3$ $\beta_2=0.7$	$\beta_1 = 0.4$ $\beta_2=0.6$	$\beta_1 = 0.5$ $\beta_2=0.5$	$\beta_1 =0.6$ $\beta_2=0.4$	$\beta_1 = 0.7$ $\beta_2=0.3$	$\beta_1 = 0.8$ $\beta_2=0.2$	$\beta_1 = 0.9$ $\beta_2=0.1$
Variaciones imputadas	1,28	1,26	1,24	1,22	1,2	1,18	1,16	1,13	1,11

Cuadro 4.
Imputaciones del salario integral para distintas combinaciones de parámetros

	Combinación de parámetros								
	$\beta_1 = 0.1$ $\beta_2=0.9$	$\beta_1 = 0.2$ $\beta_2=0.8$	$\beta_1 = 0.3$ $\beta_2=0.7$	$\beta_1 = 0.4$ $\beta_2=0.6$	$\beta_1 = 0.5$ $\beta_2=0.5$	$\beta_1 =0.6$ $\beta_2=0.4$	$\beta_1 = 0.7$ $\beta_2=0.3$	$\beta_1 = 0.8$ $\beta_2=0.2$	$\beta_1 = 0.9$ $\beta_2=0.1$
Totales Imputados	88081	86634	85186	83739	82291	80844	79397	77949	76502

Gráfica 1.
Imputaciones y dato real del total de empleados permanentes



El cálculo del coeficiente de variación para la variación del salario integral dio como resultado, 0.079 indicando que existe homogeneidad entre los datos y las imputaciones de los totales que se presentan bajo las diferentes combinaciones de parámetros, no difieren demasiado de los datos reales.

BIBLIOGRAFIA.

Roderick J.A. 1982. "Models for Nonresponse in Sample Surveys". Journal of the American Statistical Association.

Martin D. et. al. 1986. "Alternative Methods for CPS Income Imputation". Journal of the American Statistical Association.

Guerrero, V. 1991. "Análisis estadístico de series de tiempo económicas". México.

Lozano, A. 2000. "Estimación de novedades en estado de deuda". Dane.