Departamento Administrativo Nacional de Estadística

Dirección de Metodología y Producción Estadística

METODOLOGIA DE IMPUTACIÓN - 2001

MAYO 2002

CONTENIDO

ESPECIFICACIONES DE IMPUTACION

I. MARCO CONCEPTUAL.

Una gran cantidad de información, acerca de las características económicas tanto de individuos como de establecimientos industriales o países, es recopilada con fines de análisis, para entonces 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 objetivo 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 entre los resultados de diferentes análisis y reduciendo el sesgo de no respuesta.

Muchas técnicas estadísticas requieren de conjuntos de datos rectangulares o en forma de matriz y en la presencia de datos faltantes, los registros pueden restringirse 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.

II. METODOLOGÍA GENERAL DE IMPUTACIÓN DE DATOS FALTANTES

En esta metodología se utiliza la información de la muestra mensual manufacturera, 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, definida 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 establecimiento industrial y el comportamiento histórico de las variaciones dentro de cada clase industrial según la CIIU REV 3. A.C.; a partir de esta estimación se generará el dato faltante.

El cuadro-1 presenta un ejemplo, realizado para el establecimiento con número de orden (NUM_ORD)=509 y que pertenece a la clase industrial con código o CIIU3=2729,de las variaciones calculadas para las variables personal permanente, producción y ventas.

Cuadro1
Variaciones de las variables total de empleados permanentes y de la Producción de enero de 2000 a mayo de 2001

PERÍODO	TOTAL DE EMPLEADOS PERMANENTES	VARIACION DEL TOTAL DE EMPLEADOS PERMANENTES	TOTAL DE EMPLEADOS PERMANENTES POR CLASE INDUSTRIAL	VARIACION DEL TOTAL DE EMPLEADOS PERMANENTES POR CLASE INDUSTRIAL	PRODUCCION	VARIACION DE LA PRODUCCION	PRODUCCION POR CLASE INDUSTRIAL	VARIACION DE LA PRODUCCION POR CLASE INDUSTRIAL
Enero	124		425		3.033.550		15.765.529	
Febrero	124	1,00	429	1,01	3.331.605	1,10	18.321.540	1,16
Marzo	124	1,00	427	1,00	3.167.678	0,95	18.127.165	0,99
Abril	118	0,95	418	0,98	2.368.393	0,75	16.004.662	0,88
Mayo	117	0,99	417	1,00	3.099.148	1,31	17.312.161	1,08
Junio	117	1,00	401	0,96	3.140.401	1,01	18.996.862	1,10
Julio	114	0,97	401	1,00	4.265.501	1,36	19.957.135	1,05
Agosto	111	0,97	437	1,09	4.396.068	1,03	20.473.469	1,03
Septiembre	109	0,98	435	1,00	4.762.902	1,08	21.589.968	1,05
Octubre	109	1,00	436	1,00	3.715.239	0,78	20.488.398	0,95
Noviembre	112	1,03	440	1,01	5.818.543	1,57	23.159.622	1,13
Diciembre	112	1,00	438	1,00	4.174.629	0,72	21.763.649	0,94
Enero	114	1,02	393	0,90	4.314.837	1,03	22.192.023	1,02
Febrero	112	0,98	398	1,01	5.219.292	1,21	23.572.223	1,06
Marzo	112	1,00	406	1,02	6.038.856	1,16	23.805.178	1,01
Abril	112	1,00	402	0,99	5.380.395	0,89	21.436.420	0,90
Mayo	112	1,00	406	1,01	5.601.469	1,04	25.033.372	1,17

La variación del dato que se va a imputar se obtiene en términos de la variación histórica promedio en los últimos *t-k* períodos y de la variación estacional anual en los períodos t-11 hasta t-13 en la establecimiento industrial y en la clase industrial, donde k representa los datos retrasados 2 o 3 períodos y que se utilizan para observar la evolución histórica de la serie dentro del establecimiento o dentro de la clase industrial.

El modelo para imputar la variación es:

$$Variación_t = \beta_1 Vhe_{(t-1,t-2,t-3)} + \beta_2 Vhe_{(t-11,t-12,t-13)+} \beta_3 Vha_{(t-1,t-2,t-3)} + \beta_4 Vha_{(t-11,t-12,t-13)}$$

Donde,

$Variación_t$ = Variación a estimar
Vhe = variación histórica promedio por establecimiento industrial de los períodos *t-1,t-2 y t-3* o la variación estacional anual en los períodos *t-11,t-12 y t-13*

Vha = variación histórica promedio por clase industrial de los últimos k períodos o la variación estacional anual en los períodos *t-11,t-12 y t-13*
β_i para i=1,2,3 ó 4 son coeficientes de ponderación
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 clase industrial, 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."

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 clase industrial y la autocorrelación temporal entre las variaciones de los datos de la serie histórica dentro de cada establecimiento y dentro de cada clase industrial, supuestos que se utilizan en la estructura del modelo.

Ejemplo. Para explicar la metodología, se procederá a presentar un ejemplo de imputación de la variación y del total de la variable empleados permanentes y la variación y del total de la variable producción en el mes de mayo de 2001 para el establecimiento con número de orden 509.

Utilizando la base de datos en la cual se encuentran los datos históricos por establecimiento, se calcularon las variaciones que han tenido las variables por establecimiento y por cada clase industrial en los últimos 13 períodos, luego se calculó el promedio de las variaciones de los meses marzo y abril de 2001 para conocer la evolución de la variable en los dos meses anteriores, dentro del establecimiento y dentro de la clase industrial, también se calculó el promedio de las variaciones de los meses abril y mayo de 2000 para establecer la estacionalidad de las variaciones dentro del establecimiento y dentro de la clase industrial.

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 las variables mencionadas anteriormente.

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

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

TOTAL DE EMPLEADOS PERMANENTES	VARIACION DEL TOTAL DE EMPLEADOS PERMANENTES	TOTAL DE EMPLEADOS PERMANENTES POR CLASE INDUSTRIAL	VARIACION DEL TOTAL DE EMPLEADOS PERMANENTES POR CLASE INDUSTRIAL	PRODUCCION	VARIACION DE LA PRODUCCION	PRODUCCION POR CLASE INDUSTRIAL	VARIACION DE LA PRODUCCION POR CLASE INDUSTRIAL
112	1,00	406	1,01	5.601.469	1,04	25.033.372	1,17

Las imputaciones obtenidas para la variación y para el total de la variable total de empleados permanentes 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 total de empleados permanentes, en este caso solo se observa una imputación debido a que por redondeo todas las imputaciones fueron iguales.

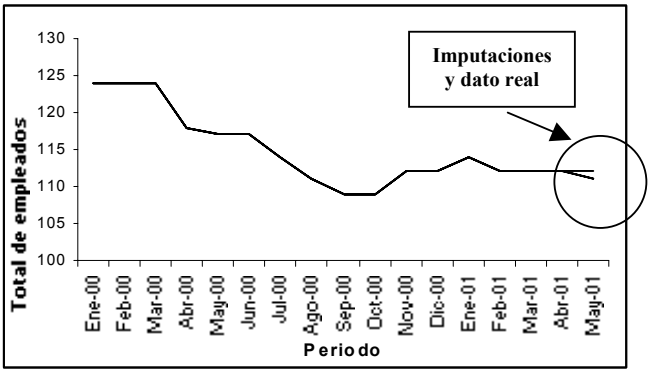
Cuadro 3.
Imputaciones de la variación del total de empleados permanentes para distintas combinaciones de parámetros

	Combinación de parámetros								
	$\beta_1 = \beta_3=0.05$ $\beta_2=\beta_4=0.45$	$\beta_1 = \beta_3=0.1$ $\beta_2=\beta_4=0.4$	$\beta_1 = \beta_3=0.15$ $\beta_2=\beta_4=0.35$	$\beta_1 = \beta_3=0.2$ $\beta_2=\beta_4=0.3$	$\beta_1 = \beta_3=0.25$ $\beta_2=\beta_4=0.25$	$\beta_1 = \beta_3=0.3$ $\beta_2=\beta_4=0.2$	$\beta_1 = \beta_3=0.35$ $\beta_2=\beta_4=0.15$	$\beta_1 = \beta_3=0.4$ $\beta_2=\beta_4=0.1$	$\beta_1 = \beta_3=0.45$ $\beta_2=\beta_4=0.05$
Variaciones imputadas	0.9933	0.9929	0.9926	0.9923	0.9920	0.9917	0.9914	0.9911	0.9908

Cuadro 4.
Imputaciones del total de empleados permanentes para distintas combinaciones de parámetros

	Combinación de parámetros								
	$\beta_1 = \beta_3=0.05$ $\beta_2=\beta_4=0.45$	$\beta_1 = \beta_3=0.1$ $\beta_2=\beta_4=0.4$	$\beta_1 = \beta_3=0.15$ $\beta_2=\beta_4=0.35$	$\beta_1 = \beta_3=0.2$ $\beta_2=\beta_4=0.3$	$\beta_1 = \beta_3=0.25$ $\beta_2=\beta_4=0.25$	$\beta_1 = \beta_3=0.3$ $\beta_2=\beta_4=0.2$	$\beta_1 = \beta_3=0.35$ $\beta_2=\beta_4=0.15$	$\beta_1 = \beta_3=0.4$ $\beta_2=\beta_4=0.1$	$\beta_1 = \beta_3=0.45$ $\beta_2=\beta_4=0.05$
Totales Imputados	111	111	111	111	111	111	111	111	111

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



Los cuadros 5 y 6 presentan respectivamente las imputaciones de la variación y del total para la variable PRODUCCION; la imputación más grande y más pequeña y el dato real del total de producción se muestran en la gráfica-2

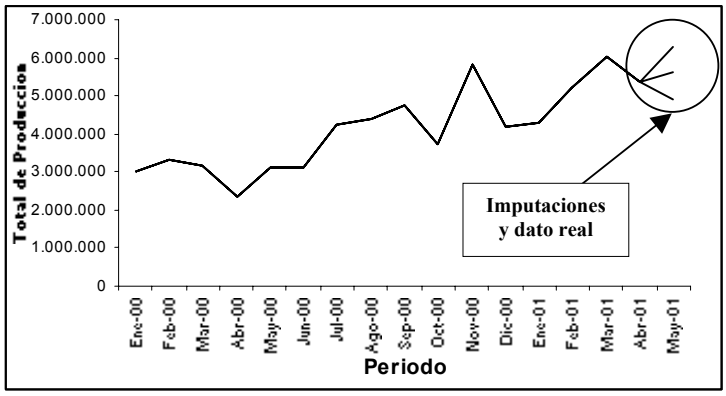
Cuadro 5.
Imputaciones de las variaciones del total de producción para distintas combinaciones de parámetros

	Combinación de parámetros								
	$\beta_1 = \beta_3=0.05$ $\beta_2=\beta_4=0.45$	$\beta_1 = \beta_3=0.1$ $\beta_2=\beta_4=0.4$	$\beta_1 = \beta_3=0.15$ $\beta_2=\beta_4=0.35$	$\beta_1 = \beta_3=0.2$ $\beta_2=\beta_4=0.3$	$\beta_1 = \beta_3=0.25$ $\beta_2=\beta_4=0.25$	$\beta_1 = \beta_3=0.3$ $\beta_2=\beta_4=0.2$	$\beta_1 = \beta_3=0.35$ $\beta_2=\beta_4=0.15$	$\beta_1 = \beta_3=0.4$ $\beta_2=\beta_4=0.1$	$\beta_1 = \beta_3=0.45$ $\beta_2=\beta_4=0.05$
Variaciones imputadas	**0,97**	**0,96**	**0,96**	**0,95**	**0,94**	**0,93**	**0,93**	**0,92**	**0,91**

Cuadro 6.
Imputaciones del total de producción para distintas combinaciones de parámetros

	Combinación de parámetros								
	$\beta_1 = \beta_3=0.05$ $\beta_2=\beta_4=0.45$	$\beta_1 = \beta_3=0.1$ $\beta_2=\beta_4=0.4$	$\beta_1 = \beta_3=0.15$ $\beta_2=\beta_4=0.35$	$\beta_1 = \beta_3=0.2$ $\beta_2=\beta_4=0.3$	$\beta_1 = \beta_3=0.25$ $\beta_2=\beta_4=0.25$	$\beta_1 = \beta_3=0.3$ $\beta_2=\beta_4=0.2$	$\beta_1 = \beta_3=0.35$ $\beta_2=\beta_4=0.15$	$\beta_1 = \beta_3=0.4$ $\beta_2=\beta_4=0.1$	$\beta_1 = \beta_3=0.45$ $\beta_2=\beta_4=0.05$
Totales Imputados	**5.231.285**	**5.189.393**	**5.147.501**	**5.105.609**	**5.063.717**	**5.021.825**	**4.979.933**	**4.938.041**	**4.896.149**

Gráfica 2.
Imputaciones mínima, máxima y dato real del total de producción



El cálculo de los coeficientes de variación para las dos variables dio como resultado: para la variable total de empleados permanentes un coeficiente de variación igual a 0.18 y para la variable Producción un coeficiente de variación igual a 0.21 indicando con esto que los datos no son muy homogéneos pero aún sin el cumplimiento de este supuesto 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.