Badler, Clara E.
Alsina, Sara M. [1]
Puigsubirá, Cristina R. [1]
Vitelleschi, María S.[1]
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística (IITAE)

UTILIZACIÓN DE METODOLOGÍA PARA EL TRATAMIENTO DE INFORMACIÓN FALTANTE Y/O CONFUSA EN EL DIAGNÓSTICO DE LA DESOCUPACIÓN[2]

1. INTRODUCCIÓN

La desocupación, al igual que otros fenómenos socio-económicos, es evaluada a través de indicadores como tasas, cuyos valores son obtenidos a partir de información relevada.

Frecuentemente se construyen tasas a partir de información disponible que no fue relevada con los requerimientos del indicador, pudiéndose obtener valores que no brindan una medición correcta del fenómeno a evaluar.

Este problema se presenta a partir de diferentes situaciones. En particular, ante la incorrecta o inapropiada clasificación de los individuos o unidades en las categorías de las variables observadas o a la excesiva estratificación de los mismos con el objeto de evaluar a subgrupos muy específicos sin un diseño que lo haya previsto.

En este trabajo, a partir de información de la Encuesta Permanente de Hogares (EPH), se presenta un análisis de la incidencia del estado ocupacional de la subpoblación de beneficiarios de Planes Jefas y Jefes de Hogares Desocupados (PJJ), en la tasa de desocupación total para el Gran Rosario y en tasas de desocupación específicas correspondientes a subgrupos determinados por categorías de variables de interés.

El estado ocupacional que la secuencia de la EPH asigna a los poseedores de PJJ es el de "ocupados", lo cual produce una disminución en la tasa de desocupación calculada a partir de dicha fuente. Este indicador en la actualidad se proporciona oficialmente incluyendo y excluyendo a dicha subpoblación.

---

[1] Docente-investigador e Investigador del Consejo de Investigaciones de la Universidad Nacional de Rosario.

[2] Proyecto SPU-UNR 0044 "La información estadística como base para el diagnóstico de la desocupación en Rosario".

A partir de la aplicación de metodología para tratamiento de información faltante ha sido considerado el estado ocupacional de los beneficiarios de PJJ como información confusa, reasignándoles la misma con técnicas de imputación a partir de la consideración de variables auxiliares, que contribuyen al acercamiento del conocimiento del verdadero estado ocupacional.

Para categorías específicas de ciertas variables se evalúan los cambios en la tasa de desocupación debido a la reasignación del estado ocupacional de los poseedores de los PJJ.

La propuesta se formula como un aporte al incremento de la calidad de la información que es utilizada para la toma de decisiones en el ámbito económico-social.

2. MATERIAL

La información proviene de la Base Usuaria Ampliada y de la Base de Jefas/Jefes de Hogar de la EPH, onda mayo 2003, correspondiente al Aglomerado Gran Rosario.

La variable "Estado Ocupacional", presenta las categorías: (1) ocupado, (2) desocupado y (3) inactivo.

Las covariables utilizadas son: sexo, edad, relación de parentesco, estado civil, asistencia a establecimiento escolar e ingreso total.

Para el cálculo de la tasa de desocupación por categoría de variable, las siguientes fueron recategorizadas:

- Escolaridad: (1) hasta primario incompleto;
  (2) primario completo, secundario incompleto;
  (3) secundario completo, superior incompleto, universitario incompleto;
  (4) superior completo, universitario completo.
- Ingreso: (.) sin ingreso, ingresos parciales o ns/nr;
  (1) decíl del ingreso individual igual a 1, 2 o 3;
  (2) decíl del ingreso individual igual a 4, 5, 6, o 7;
  (3) decíl del ingreso individual igual a 8, 9 y 10.
- Edad: (0) menores de 16 años;
  (1) mayores de 15 y menores de 25 años;

(2) mayores de 24 y menores de 46;

(3) mayores de 45 y menores de 66;

(4) mayores de 65.

- Rama de actividad: (0) no corresponde;
  (1) construcción;
  (2) manufactura;
  (3) servicios comerciales y de transporte;
  (4) intermediación financiera;
  (5) administración pública y defensa;
  (6) instrucción pública y servicios de salud;
  (7) otras actividades de servicios.

3. METODOLOGÍA

3.1 E·squema de pérdida monótono

Una base de datos multivariada con variables Y_1, Y_2,...,Y_p, presenta un esquema monótono de pérdida cuando se observa que, si la variable Y_j presenta pérdidas para un individuo en particular, todas las subsiguientes variables Y_k , k>j, presentan faltas para ese individuo.

3.2 Método de Imputación por Regresión

Es un método paramétrico de imputación que utiliza modelos de regresión para bases de datos con esquemas monótonos de pérdida.

Consiste en ajustar un modelo de regresión para cada variable con valores faltantes, utilizando como covariables las variables previas en el esquema monótono. Basándose en el modelo resultante, un nuevo modelo de regresión se simula y es utilizado para imputar los valores faltantes de cada variable. Como el conjunto de datos presenta un esquema monótono de pérdida, el proceso se repite secuencialmente para las variables con valores faltantes.

Para una variable Y_j con valores faltantes, el modelo:

$$Y_j = \beta_o + \beta_1 Y_1 + \beta_2 Y_2 + ...+ \beta_{(j-1)} Y_{(j-1)}$$

se ajusta con sólo los valores observados de las variables $Y_1, Y_2, ..., Y_{j-1}$.

Para cada imputación, los nuevos parámetros ($\beta_{*0}, \beta_{*1}, ..., \beta_{*(j-1)}$) y σ^2_{*J} son simulados a partir de $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{j-1}), \sigma_j^2$ y $\mathbf{V_j}$. La variancia obtenida es $\sigma^2_{*j} = \hat{\sigma_j^2}(n_j - j)/g$, donde g es una variable aleatoria con distribución $\chi^2_{(nj-j)}$ y n_j es el número de observaciones completas para Y_j.

Los coeficientes de regresión son obtenidos como $\beta_* = \hat{\beta} + \sigma_{*j}\, \mathbf{V}'_{\mathbf{hj}}\, \mathbf{Z}$, donde $\mathbf{V}'_{hj}$ es la matriz triangular superior en la descomposición de Cholesky, $\mathbf{V}_j = \mathbf{V}'_{hj}\, \mathbf{V}_{hj}$, y $\mathbf{Z}$ es un vector de j variables aleatorias normales independientes.

Los valores perdidos son reemplazados entonces por:

$$\beta_{*o} + \beta_{*1}\, y_1 + \beta_{*2}\, y_2 + ... + \beta_{*(j-1)}\, y_{(j-1)} + z_i\, \sigma_{*J}$$

donde $y_1, y_2, ..., y_{(j-1)}$ son los valores de las covariables de las primeras (j-1) variables y z_i es una simulación de un desvío normal.

Este método de imputación por regresión se implementa mediante un procedimiento del programa SAS denominado PROC MI.

Se aplica este método bajo el supuesto que los datos son perdidos al azar (MAR), o sea que la probabilidad de que una observación sea faltante depende de los valores observados pero no de los faltantes.

4. RESULTADOS

Todos los individuos que declaran disponer de un PJJ en su ocupación principal en el formulario individual, tienen asignado en la variable correspondiente al estado ocupacional la categoría "ocupado".

Dentro de este grupo, como forma de acercamiento al real estado ocupacional se define un subgrupo con las siguientes características:

-Declaran poseer PJJ en su ocupación principal.

-El establecimiento en el cual se desempeñan corresponde el sector público.

-Tienen una única ocupación.

-El monto del ingreso total es de $150.

Se considera que el subgrupo definido tiene información confusa en la variable ”estado ocupacional” y que los individuos del mismo están incorrectamente clasificados en las categorías de dicha variable, y a través de la metodología presentada se los reclasifica en las categorías desocupado e inactivo.

Mediante el método de imputación basada en un modelo de regresión para esquema monótono de pérdida y considerando las variables auxiliares: edad, relación de parentesco, asistencia a establecimiento escolar e ingreso total a través del PROC MI de SAS, se reasigna el estado ocupacional al subgrupo definido anteriormente (Tabla 1).

Tabla 1: Reclasificación de los beneficiarios de PJJ en categorías de “Estado Ocupacional”.

Estado Ocupacional	Base original	Base imputada
0	1	1
Ocupados	574	525
Desocupados	129	164
Inactivos	919	933
Total	1623	1623

El 71,43% de los individuos del subgrupo definido, que originalmente eran considerados ocupados se reclasifican en desocupados y el 28,57% en inactivos.

A partir de la nueva distribución de frecuencias de la variable “Estado Ocupacional” se calcula la tasa de desocupación para el Gran Rosario, que es comparada con la obtenida de la información proporcionada originalmente por la EPH (Tabla 2).

Tabla 2: Tasas de desocupación original y corregida.

Información	Tasa de desocupación
Original	0.1835
Corregida	0.2380

Al aplicar la metodología propuesta se recalcula la tasa de desocupación para cada categoría de las variables de interés (Tablas 3.....7).

Tabla 3: Tasas de desocupación original y corregida para la variable sexo

Sexo	Base Original	Base Imputada
Hombres	0.1856	0.2039
Mujeres	0.1806	0.2857

Tabla 4: Tasas de desocupación original y corregida para la variable edad

Edad	Base Original	Base Imputada
1	0.4203	0.4667
2	0.1131	0.1921
3	0.1509	0.1770
4	0.0625	0.0625

Tabla 5: Tasas de desocupación original y corregida para la variable escolaridad

Escolaridad	Base Original	Base Imputada
1	0.1617	0.2985
2	0.1961	0.2616
3	0.2202	0.2511
4	0.073	0.0830

Tabla 6: Tasas de desocupación original y corregida para la variable ingreso

Ingreso	Base Original	Base Imputada
1	0.097	0.40
2	0.02793	0.02793
3	0	0

Tabla 7: Tasas de desocupación original y corregida para la variable rama de actividad

Rama de actividad	Base Original	Base Imputada
1	0.3238	0.3238
2	0.6071	0.6071
3	0.1934	0.1934
4	0.085	0.085
5	0.062	0.2370
6	0.1143	0.1827

Se observa disparidad en las variaciones entre las tasas calculadas con la base original y con la que surge al aplicar la metodología propuesta.

5. DISCUSIÓN

- La metodología para el tratamiento de la información faltante constituye una herramienta para el abordaje de bases de datos provenientes de encuestas con información confusa, contribuyendo a incrementar la calidad de la información utilizada en la construcción de indicadores que reflejen la realidad económico-social.

- Se proporcionan elementos a incorporar en el debate sobre la incidencia de los PJJ en la situación ocupacional general.

- Debe señalarse que hay que actuar con cautela al calcular la tasa de desocupación por categorías de variable ya que existen limitantes sobre todo las relacionadas con el número de observaciones.

6. REFERENCIAS

Allison, P.D.. (2000). *"Multiple imputation for missing data: A cautionary tale". Sociological Methods and Research, vol. 28, pp. 301-309.*

Little, R. J. and D. B. Rubin. (2002). "*Statistical Analysis with Missing Data*". Second Edition. John Wiley & Sons, New York.

Rubin, D.. (1987). "*Multiple imputation for nonresponse in surveys*". John Wiley & Sons. , New York.

SAS® Institute Inc., Statistics and Operations Research. "What´s new in data analysis. multiple imputation for missing data". 2002. http://www.sas.com/rnd/app/da/new/ dami.htm. (Mayo 2002)

Schafer, J. L.. (1997). *"Analyses of Incomplete Multivariate Data"*. Chapman & Hall. Londres.

Yuan,Y.C..(2001)."Multiple imputation for missing data: concepts and new development". SUGI Proceedings. http://www.ats.ucla.edu/stat/sas/library/ default.htm, (Octubre 2001).