Badler, Clara E.
Alsina, Sara M. [1]
Puigsubirá, Cristina B. [1]
Vitelleschi, María S.[1]

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística (IITAE)

INFORMACIÓN FALTANTE Y/O CONFUSA EN VARIABLES DISCRETAS DE LA ENCUESTA PERMANENTE DE HOGARES (EPH)

1. INTRODUCCIÓN

Parte de la información relevada en la EPH se refiere a variables discretas que dan origen a tablas de contingencias RxC. Cuando se presenta información faltante y/o confusa, en dichas tablas la información resulta incompletamente clasificada. Es frecuente el problema de estimar las probabilidades de celda en estas situaciones.

Existen diferentes métodos para las estimaciones máximo verosímiles de las probabilidades de celda en tablas de contingencia RxC incompletas cuando las variables fila y columna se distribuyen conjuntamente multinomial y el mecanismo de pérdida es ignorable. En estos casos tanto el algoritmo EM como el de Newton-Raphson requieren modificaciones de los programas estadísticos computacionales existentes, mientras que S. R. Lipsitz y colaboradores propusieron un procedimiento basado en la conexión entre las verosimilitudes de las distribuciones multinomial y de Poisson, mostrando que dichas estimaciones pueden ser obtenidas a partir de cualquier procedimiento de modelos lineales generalizados, como PROC GENMOD de SAS.

En este trabajo se utiliza dicha metodología para obtener las estimaciones máximo verosímiles de las probabilidades de celda de tablas de contingencia RxC que surgen de variables discretas de la EPH relacionadas con el estado ocupacional relevadas para todos los aglomerados del país, con información faltante y/o confusa.

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

2. MATERIAL

La información proviene de la onda mayo 2003 de la EPH para todos los aglomerados del país, utilizando la Base Usuaria Ampliada y la Base Jefas y Jefes de Hogar.

Variables utilizadas en el análisis:

- "Estado Ocupacional":1- Ocupado
  2- Desocupado
  3- Inactivo
  9- Información Confusa
- "Nivel Educacional": 1- Primario Completo e incompleto y enseñanza media incompleta
  2- Enseñanza media completa y enseñanzas superior y universitaria incompleta.
  3- Enseñanzas superior y universitaria completa.
  9- Pérdidas simuladas.

3. METODOLOGÍA

Se usa el procedimiento propuesto por Lipsitz y colaboradores para estimar las probabilidades de celda (p_{jk}) en tablas de contingencias incompletas, es decir aquellas en que las variables fila o columna están perdidas para algunos de los individuos.

Se considera el caso de tablas bidimensionales en que las pérdidas son al azar o completamente al azar y el mecanismo que las produce es ignorable.

Se aplica el procedimiento que establece la conexión entre las verosimilitudes de las distribuciones multinomial y la de Poisson, para mostrar que las estimaciones máximo verosímiles de las probabilidades de celda de tablas de contingencia RxC con información confusa o perdida, pueden ser obtenidas con un procedimiento para modelos lineales generalizados.

La función de distribución conjunta de "datos completos", para la i-ésima observación, bajo el supuesto que las pérdidas son al azar (MAR) es:

siendo: Y_1 e Y_2: variables aleatorias discretas; **R**$_1$ y **R**$_2$: variables aleatorias indicadoras de pérdida; **p** y **ϕ** parámetros y $f(r_{i1}, r_{i2} / y_{i1}, y_{i2}, \phi)$ mecanismo de pérdida.

Para estimar **p,** se obtiene la siguiente expresión para la función de verosimilitud, en el caso de N individuos independientes:

$$L(p) = \prod_{i=1}^{N} \left[f(y_{i1}, y_{i2} / p)^{r_{i1} r_{i2}} \; f(y_{i1} / p)^{r_{i1}(1-r_{i2})} \; f(y_{i2} / p)^{(1-r_{i1}) r_{i2}} \right]$$

La que se puede reescribir como:

$$L(p) = \left[\prod_{j=1}^{R} \prod_{k=1}^{C} p^{u_{jk}}{}_{jk} \right] \left[\prod_{j=1}^{R} p^{w_{j+}}{}_{j+} \right] \left[\prod_{k=1}^{C} p^{z_{+k}}{}_{+k} \right]$$

siendo el producto de las contribuciones a la verosimilitud de los casos completos (u_{jk}), de los individuos sólo observados en Y_{i1} (w_{j+}) y de los individuos sólo observados en Y_{i2} (z_{+k}).

Dado que: una verosimilitud multinomial puede ser escrita como una de Poisson, que la verosimilitud multinomial es función de las probabilidades p_{jk} mientras la de Poisson es función de las frecuencias esperadas de celdas y que u_{jk}, w_{j+} y z_{+k} son variables aleatorias independientes de Poisson, la L(p) puede ser reescrita en función de las frecuencias esperadas de celdas si se cumple que:

$$p_{++} = \sum_{j=1}^{R}\sum_{k=1}^{C} p_{jk} = 1 \quad o\ sea \quad p_{RC} = 1 - \sum_{jk \neq RC} p_{jk}$$

Se puede maximizar L(p) usando un modelo lineal generalizado de Poisson con expresión:

$$E(\mathbf{f}) = \mathbf{Xp} + \mathbf{g}$$

donde: **f**: vector de frecuencias de celda

X: matriz de diseño

p: vector de probabilidades de celda

g: vector de totales marginales usado como variable offset.

Las estimaciones de **p** se obtienen mediante una macro en SAS (MISSRC), que construye **X** y **g** y luego implementa el PROC GENMOD. **4. RESULTADOS**

En el contexto del análisis de la relación entre ocupación y educación, se construye la tabla de contingencia bidimensional con las variables "Estado Ocupacional y "Nivel Educacional" para el subgrupo "Individuos de 18 años y más que asisten o asistieron a la escuela".

Para la aplicación de la metodología propuesta:

- En la variable "Estado Ocupacional" se consideran valores confusos a los correspondientes a aquellos individuos que declaran ser ocupados y disponer de un Plan Jefas y Jefes en su ocupación principal.

- Se generan pérdidas completamente al azar (MCAR) en la variable "Nivel Educacional" en un porcentaje de 25% del total de observaciones que resultan distribuidas aleatoriamente en las respectivas categorías de la variable.

De esta manera se obtiene una tabla de contingencia con valores completamente clasificados y otros sólo parcialmente clasificados (Tabla 1).

Tabla 1: Cantidad de individuos con 18 años o más que asisten o asistieron a la escuela según “Estado Ocupacional” y “Nivel educacional”.

Niv. Educ. / Est. Ocup	1	2	3	9
1	4304	7185	2817	4620
2	769	1460	252	828
3	5025	5559	967	3845
9	888	760	79	0

Para estimar las probabilidades de celda se aplica la metodología utilizando la macro de SAS (MISSRC) que trabaja bajo el supuesto de que las pérdidas obedecen a un mecanismo ignorable.

La macro crea una base de datos que contiene la matriz de diseño construida para el modelo lineal generalizado de Poisson, el vector de los totales marginales usado como una variable offset y las frecuencias disponibles de las celdas (Tabla 2).

Tabla 2: Frecuencias observadas, matriz de diseño y vector de totales marginales.

COUNT	P11	P12	P13	P21	P22	P23	P31	P32	OFFSET
4304	28338	0	0	0	0	0	0	0	0
7185	0	28338	0	0	0	0	0	0	0
2817	0	0	28338	0	0	0	0	0	0
769	0	0	0	28338	0	0	0	0	0
1460	0	0	0	0	28338	0	0	0	0
252	0	0	0	0	0	28338	0	0	0
5025	0	0	0	0	0	0	28338	0	0
5559	0	0	0	0	0	0	0	28338	0
967	-28338	-28338	-28338	-28338	-28338	-28338	-28338	-28338	28338
888	1727	0	0	1727	0	0	1727	0	0
760	0	1727	0	0	1727	0	0	1727	0
79	-1727	-1727	0	-1727	-1727	0	-1727	-1727	1727
4620	9293	9293	9293	0	0	0	0	0	0
828	0	0	0	9293	9293	9293	0	0	0
3845	-9293	-9293	-9293	-9293	-9293	-9293	0	0	9293

Mediante el PROC GENMOD incorporado en la macro se obtiene la estimación puntual y por intervalo de los parámetros del modelo. En la tabla 3 se presentan las estimaciones de las probabilidades de celda y las correspondientes frecuencias ajustadas.

Tabla 3: Estimaciones máximo verosímiles de las probabilidades de celda, errores estandares asociados y frecuencias ajustadas, según categorías de las variables "Estado Ocupacional" (EST) y "Nivel Educacional" (NEDU).

EST	NEDU	PARM	ESTIMATE	STDERR	FITTED
1	1	P11	0.15528	.002094793	6111.43
1	2	P12	0.25105	.002434207	9880.64
1	3	P13	0.09527	.001675418	3749.51
2	1	P21	0.02795	.000952822	1100.15
2	2	P22	0.05138	.001213095	2022.38
2	3	P23	0.00858	.000532183	337.79
3	1	P31	0.18232	.002192685	7175.62
3	2	P32	0.19530	.002230534	7686.53
3	3	P33*	0.03288	.001031390	1293.95

Se destaca con * la probabilidad correspondiente a la p_{33}, calculada como: $\hat{p}_{JK} = 1 - \sum_{jk \neq JK} \hat{p}_{JK}$.

5. DISCUSIÓN

- Se propone una solución para la estimación máximo verosímil de las probabilidades de celda en tablas de contingencia RxC con información faltante y/o confusa mediante la aplicación de modelos lineales generalizados.

- La ventaja de este procedimiento es la posibilidad de su aplicación sin requerir programación adicional.

- En este trabajo se ha implementado la metodología propuesta mediante una macro disponible en SAS. También sería posible realizarlo con otros programas como GLIM o SPLUS.

- El procedimiento puede ser generalizado para tablas de contingencia de mayor dimensión con resultados análogos.

6. REFERENCIAS

Friendly, M.(2000).”Note on ‘Obtaining the maximum likelihood estimates in incomplete RxC contingency tables…’”. Journal of Computational and Graphical Statistics, 9(1), 158-166.

Lipsitz, S.; Parzen, M. and Molenberghs, G.. (1998). “Obtaining the Maximum likelihood estimates in incomplete RxC contingency Tables using a Poisson generalized linear model”. Journal of Computational and Graphical Statistics, 7, 356-376.

Little, R. And Rubin, D. (2002). “*Statistical Analysis with Missing Data*”. Second Edition. John Wiley and Sons, New York.