INFLUENCIA DEL MECANISMO, DE LA ESTRUCTURA DE LOS DATOS Y DEL TRATAMIENTO DE LA INFORMACIÓN INCOMPLETA EN EL ANALISIS ESTADÍSTICO DE LA EPH

Badler, Clara; Alsina, Sara; Beltrán, Celina; Puigsubirá, Cristina; Vitelleschi, Ma. Susana

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística, Fac. Cs. Económicas y Estadística. Universidad Nacional de Rosario

INTRODUCCIÓN

- El mecanismo que produce la pérdida
- El tratamiento para la falta de información

pueden afectar → La estructura de los datos. → modificada alteran → Los resultados de la técnica propuesta.

Los procedimientos de simulación de pérdidas y el uso posterior de técnicas de análisis multivariado permiten describir e interpretar un determinado fenómeno, considerando diferentes situaciones.

Para comparar los resultados se aplicó la técnica de Componentes Principales a distintos conjuntos de datos → Original; Casos completos; Con datos imputados

METODOLOGÍA

MECANISMOS DE PÉRDIDA

MCAR
Cuando la probabilidad de respuesta es independiente de las variables observadas completamente y de las observadas en forma parcial.

MAR
Cuando la probabilidad de respuesta depende de las variables observadas completamente pero no de las observadas en forma parcial.

Cuando la probabilidad de respuesta depende de las variables observadas completamente y posiblemente de las observadas en forma parcial, el mecanismo no es MCAR ni MAR.

Para la evaluación del supuesto MCAR se utiliza un test propuesto por Little, aplicado a un esquema monótono con tres variables, en el que se observa:

- Y_1 en forma completa (n unidades)
- Y_2 en n_2 (<n) unidades
- Y_3 en n_3 ($<n_2$) unidades

$$d^2 = \frac{SSB_1}{MST_1} + \frac{SSB_{2.1}}{MST_{2.1}} = \frac{(n-1)\,2\,F_1}{(n-3)+2\,F_1} + \frac{(n_2-1)\,F_{2.1}}{(n_2-2)+F_{2.1}} \sim \chi^2_3$$

- SSB_1, MST_1 y F_1 son la suma de cuadrados entre grupos, el cuadrado medio total y la estadística F del análisis de la variancia de Y_1 sobre el esquema de pérdida
- $SSB_{2.1}$, $MST_{2.1}$ y $F_{2.1}$ son la suma de cuadrados entre grupos, el cuadrado medio total y la estadística F del análisis de covariancia de la variable Y_2 sobre los grupos determinados por los esquemas donde Y_2 es observada, ajustando por Y_1

ESTUDIO DE LA ESTRUCTURA DE LOS DATOS

- Se divide el conjunto de datos originales en grupos con respecto a una determinada variable (Y_j). En este caso, la variable presenta pérdida de información simulada.
- Se utiliza como punto de corte para dicha división los percentiles de la variable Y_j.
- Se calcula la correlación entre las variables que presentan información incompleta en cada uno de los subconjuntos de datos resultantes de la división y la matriz de correlaciones para el conjunto completo de datos.
- Se analiza la tendencia que presentan las correlaciones en los distintos grupos.
- Se investiga de qué forma se ven afectados los resultados obtenidos de la aplicación de técnicas estadísticas basadas en las correlaciones.

TRATAMIENTO DE LA INFORMACION FALTANTE

CASOS COMPLETOS:
Consiste en utilizar sólo las unidades con información completa en todas las variables. Es adecuado bajo el supuesto MCAR.

IMPUTACIÓN POR EL MÉTODO DE BUCK:
Asigna al valor perdido el obtenido como predicción de la regresión lineal sobre las variables observadas en esa unidad:

$$\hat{y}_{ij.per} = \hat{\beta}' \mathbf{Y}_{obs.j}$$

- $y_{ij.per}$ valor faltante de la variable Y_j en la i-ésima unidad
- $\mathbf{Y}_{obs.i}$ vector de los valores observados en dicha unidad
- $\hat{\beta}$ vector de coeficientes estimados de regresión.

Realiza el supuesto MCAR, sin embargo los estimadores obtenidos pueden ser consistentes aún cuando el mecanismo sea MAR.

COMPONENTES PRINCIPALES

Transforma el conjunto de datos en uno de menor dimensión a través de nuevas variables que expliquen el máximo de la variación de las variables originales

$CP_j = a_{1j}\,Y_1 + a_{2j}\,Y_2 + \ldots + a_{pj}\,Y_p \qquad j=1,2,\ldots,$

$Var(CP_j) = \mathbf{a}'_j\,\Sigma\,\mathbf{a}_j = \lambda_j \qquad Cov(CP_j, CP_i) = \mathbf{a}'_j\,\Sigma\,\mathbf{a}_i \qquad j \neq i$

Σ = matriz de variancias y covariancias de las variables Y.
$\mathbf{a}_j$ = autovector normalizado asociado al j-ésimo autovalor de la matriz de correlaciones.

RESULTADOS

Bases de datos de la EPH:
- Onda Octubre 1997
- Onda Octubre 1998
Aglomerado Gran Rosario

→ **Sub bases:** Personas desocupadas con ocupación anterior

→ **Variables:**
- EDAD (años cumplidos a la fecha del relevamiento)
- TTOA: Tiempo transcurrido desde que dejó la ocupación anterior (en días)
- TBE: Tiempo de búsqueda de empleo (en días)

SIMULACIÓN DE PERDIDA DE INFORMACIÓN

Mediante el programa SAS se generan pérdidas para obtener un esquema monótono en el cual la variable EDAD se observa en forma completa, TTOA presenta 10% de pérdida y TBE un 20%. Las simulaciones se realizan bajo distintos mecanismos de pérdida:

MCAR
Se seleccionan en forma aleatoria las observaciones que tendrán información faltante en las variables TBE y TTOA.

MAR
Se seleccionan al azar observaciones consideradas como "potenciales pérdidas" en TBE y TTOA. Luego se consideran como perdidas aquellas en que la EDAD supera los 40 años.

NO ALEATORIO
- Se seleccionan los mayores valores de TTOA y se generan pérdidas en TTOA y TBE. Luego, de los restantes valores de TTOA, se seleccionan los mayores y se generan pérdidas en TBE.
- Se seleccionan los mayores valores de TTOA y se generan pérdidas en TTOA y TBE. Luego, de los restantes valores de TTOA, se seleccionan los mayores y se generan pérdidas en TBE.

Proyectos: Pict Nº 0200095-01996 de la ANPCyT y PID Nº 19/E045 de la SECyT

Cuadro 1: Test para evaluar el supuesto MCAR

ONDA	MECANISMO DE PERDIDA	F_1	$F_{2.1}$	d^2 (GL=3)	PROB	DECISIÓN*
Octubre 1997 (N=171)	MCAR	1.1636	0.639	2.96	0.40	No se rechaza
	MAR	35.9433	0.031	50.975	$4.95E^{-11}$	Se rechaza
	Valores máx.	0.7435	182.77	85.02	$2.57E^{-18}$	Se rechaza
	Valores mín.	0.01	8.13	7.9	0.04	Se rechaza
Octubre 1998 (N=142)	MCAR	0.0518	0.456	0.56	0.90	No se rechaza
	MAR	25.65	2.346	40.33	$9.06E^{-9}$	Se rechaza
	Valores máx.	2.82	400.28	102.09	$5.52E^{-22}$	Se rechaza
	Valores mín.	0.1035	0.759	0.97	0.81	No se rechaza

*α=0.05

- Para la onda Octubre de 1997, en todo los mecanismos no MCAR, se rechaza la hipótesis
- Para la onda Octubre de 1998, en el mecanismo no aleatorio por el cual se pierde información en unidades con valores bajos de TTOA, no se rechaza la hipótesis.

CAMBIOS DETECTADOS EN LA CORRELACIÓN TTOA*TBE

Cuadro 2: Cambios producidos en la correlación TTOA*TBE. Onda Octubre de 1997

Conjunto de datos	Correlación TTOA*TBE
Datos originales	0.43
Mec. No aleatorio Valores Máx. (Casos Completos)	0.85
Mec. No aleatorio Valores Máx. (Imputación Buck)	0.92

Cuadro 3: Cambios producidos en la correlación TTOA*TBE. Onda Octubre de 1998

Conjunto de datos	Correlación TTOA*TBE
Datos originales	0.02
Mec. No aleatorio Valores Máx. (Casos Completos)	0.78
Mec. No aleatorio Valores Máx. (Imputación Buck)	0.93

La correlación entre las variables que presentan pérdidas se ve sobrestimada al trabajar con los casos completos como al imputar por el método de Buck.

ESTRUCTURA DE LOS DATOS

Se divide la muestra en grupos definidos por los percentiles 20, 40, 60 y 80 de la variable TTOA

Gráfico 1: Correlación TTOA*TBE según grupo de TTOA . Onda Octubre de 1997

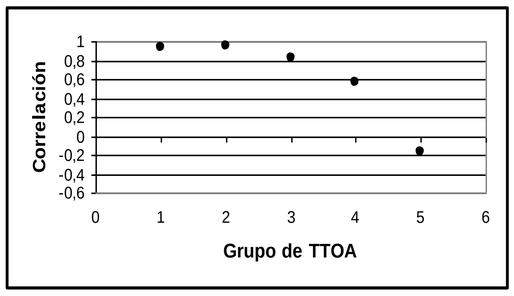
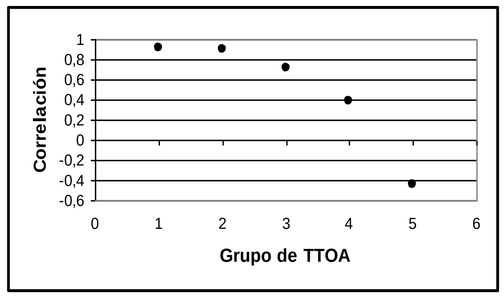


Gráfico 2: Correlación TTOA*TBE según grupo de TTOA . Onda Octubre de 1998



La correlación entre las variables TTOA y TBE disminuye al aumentar el valor de la variable TTOA. → Al perder los valores más altos de TTOA, se utiliza información de las unidades que presentan una correlación mayor que la de la muestra original. → Sobrestimación de la correlación.

APLICACIÓN DE COMPONENTES PRINCIPALES

Se presentan los resultados para los mecanismos MCAR y NO ALEATORIO (Valores máx), ya que en estas situaciones los mismos presentan diferencias.

MECANISMO MCAR

Cuadro 4: Coeficientes y porcentaje de variancia explicada de las dos primeras componentes, onda Octubre 1997.

Variable	COEFICIENTES					
	DATOS ORIGINALES		METODO DE BUCK		CASOS COMPLETOS	
	CP1	CP2	CP1	CP2	CP1	CP2
EDAD	0.37	0.92	0.36	0.93	0.35	0.93
TBE	0.67	-0.15	0.67	-0.17	0.67	-0.17
TTOA	0.64	-0.37	0.65	-0.34	0.65	-0.32
% de variancia explicada	81.2		80.2		80.1	

Cuadro 5: Coeficientes y porcentaje de variancia explicada de las dos primeras componentes, onda Octubre 1998.

Variable	COEFICIENTES					
	DATOS ORIGINALES		METODO DE BUCK		CASOS COMPLETOS	
	CP1	CP2	CP1	CP2	CP1	CP2
EDAD	0.69	-0.05	0.69	-0.03	0.70	-0.03
TBE	0.40	0.86	0.43	0.81	0.39	0.85
TTOA	0.60	-0.51	0.58	-0.58	0.60	-0.52
% de Variancia explicada	73.3		72.5		72.4	

MECANISMO NO ALEATORIO (Valores máx.)

Cuadro 6: Coeficientes y porcentaje de variancia explicada de las dos primeras componentes, onda Octubre 1997.

Variable	COEFICIENTES					
	DATOS ORIGINALES		METODO DE BUCK		CASOS COMPLETOS	
	CP1	CP2	CP1	CP2	CP1	CP2
EDAD	0.37	0.92	0.24	0.97	0.26	0.97
TBE	0.67	-0.15	0.69	-0.15	0.69	-0.15
TTOA	0.64	-0.37	0.68	-0.19	0.68	-0.21
% de Variancia explicada	81.2		95.0		97.5	

Cuadro 7: Coeficientes y porcentaje de variancia explicada de las dos primeras componentes, onda Octubre 1998.

Variable	COEFICIENTES					
	DATOS ORIGINALES		METODO DE BUCK		CASOS COMPLETOS	
	CP1	CP2	CP1	CP2	CP1	CP2
EDAD	0.69	0.05	-0.02	0.99	0.05	0.99
TBE	0.40	-0.86	0.71	0.02	0.71	-0.01
TTOA	0.60	0.51	0.71	0.01	0.71	-0.05
% de Variancia explicada	73.3		97.8		92.7	

Se puede observar como los resultados de la aplicación de componentes principales se ven afectados cuando se ha modificado la estructura de las relaciones entre las variables.

DISCUSIÓN

Al analizar bases de datos con información incompleta debe considerarse:
- ➔ la naturaleza de las pérdidas
- ➔ los supuestos sobre las características del mecanismo que las produjo
- ➔ el tratamiento aplicado para "completar" los valores perdidos
- ➔ la estructura de los datos originales y la resultante de la aplicación del tratamiento seleccionado.