Revista Española de Salud Pública
ISSN: 1135-5727
resp@msc.es
Ministerio de Sanidad, Servicios Sociales e Igualdad
España

Barroso Utra, Isabel María; Cañizares Pérez, Mayilée; Lera Marqués, Lydia

INFLUENCIA DE LA ESTRUCTURA DE LOS DATOS EN LA SELECCIÓN DE LOS MÉTODOS DE ANÁLISIS ESTADÍSTICOS

Revista Española de Salud Pública, vol. 76, núm. 2, marzo-abril, 2002
Ministerio de Sanidad, Servicios Sociales e Igualdad
Madrid, España

Rev Esp Salud Pública 2002; 76: 95-103 N.º 2 - Marzo-Abril 2002

COLABORACIÓN ESPECIAL

INFLUENCIA DE LA ESTRUCTURA DE LOS DATOS EN LA SELECCIÓN DE LOS MÉTODOS DE ANÁLISIS ESTADÍSTICOS

Isabel María Barroso Utra (1), Mayilée Cañizares Pérez (1) y Lydia Lera Marqués (2)

(1) Instituto Nacional de Higiene, Epidemiología y Microbiología.
(2) Instituto de Cibernética, Matemática y Física.

RESUMEN

En las investigaciones médicas se encuentran datos agrupados ya sea por el diseño del estudio o la selección de la muestra. Esta estructura debe ser considerada para obtener estimaciones apropiadas de los parámetros y sus errores estándares. El presente trabajo es de naturaleza metodológica y se destina a ilustrar métodos para estimar parámetros poblacionales y modelos de regresión con datos agrupados. Para ello se utilizan nueve variables de la *I Encuesta Nacional de Factores de Riesgo y Actividades Preventivas,* realizada en Cuba en 1995. La prevalencia de hipertensión arterial se sobreestima en un 15% cuando se utilizan los estimadores convencionales comparado con el análisis con pesos y el ajustado. En los modelos de regresión para el índice de masa corporal se encontró que con los procedimientos convencionales: sexo, nivel de educacional, condición de sedentarismo, tabaquismo, tensión diastólica y sistólica resultaron significativas. Sin embargo, con el método que considera la estructura de conglomerados dejaron de ser significativas el nivel educacional y la condición de sedentarismo. Al ajustar el modelo de intercepto aleatorios se encontró que el 91,3% de la variabilidad total se explica por variables individuales y el 8,7% se atribuye a unidades superiores. Al estimar parámetros poblacionales en datos con estructura de conglomerados y con desigualdad en las probabilidades de selección hay que considerar el uso de pesos muestrales y métodos de análisis que contemplen la correlación entre sujetos (potencial) de un mismo conglomerado. Al ajustar modelos de regresión no sólo importa obtener eficiencia en la estimación de los coeficientes sino que se debe considerar el enfoque (agregado o desagregado) para modelar el problema objeto de estudio.

Palabras Claves: Muestreo por conglomerados. Datos jerárquicos. Efectos aleatorios. Agregación de datos. Análisis de regresión.

ABSTRACT

The Influence of Data Structure for Selecting Statistical Analysis Methods

In medical research, data is grouped either as per the design of the study or the selection of the sample. This structure must be taken into account in order to make correct estimates of the of the parameters and standard errors involved.. This study is of the methodological type and is aimed at illustrating methods for estimating population-related methods and regression models with grouped data. For this purpose, nine variables from the *First National Risk Factor and Preventive Measure Survey* conducted in Cuba in 1995 are employed. The prevalence of high blood pressure is overestimated by 15% when the conventional estimators are used as compared with the weight-based and adjusted analysis. In the regression models for the body mass index based on the conventional procedures, sex, degree of schooling, degree of sedentariness, smoking habit, diastolic and systolic blood pressure were found to be significant. However, when the method taking into account the structure of conglomerates was employed, the degree of schooling and sedentariness ceased to be significant. When the random intercept model was adjusted, the 91.3% total variability was found to be explained by individual variables, the 8.7% variability being attributed to larger units. When estimating population-related parameters based on conglomerate-structure data involving inconsistent selection probabilities, the use of sample-related weights and analysis methods that take in the correlation among subjects (potential) for one same conglomerate. When adjusting regression models, it is not only important to efficiently estimate the coefficients, but rather the focus (aggregated or disaggregated) must be taken into account for modeling the problem under study.

Key words: Conglomerates. Databases. Random effects. Aggregated. Disaggregated.

INTRODUCCIÓN

La estructura en los datos que se presenta con mayor frecuencia en los estudios espidemiológicos y de salud pública es la *jerárquica o de conglomerados*. Por lo general, esto se debe al diseño del estudio y a la for-

Correspondencia:
Isabel María Barroso
Instituto Nacional de Higiene, Epidemiología y Microbiología
Infanta 1158 entre Clavel y Llinas.
La Habana, Cuba.
Correo electrónico: ibarroso@inhem.sld.cu

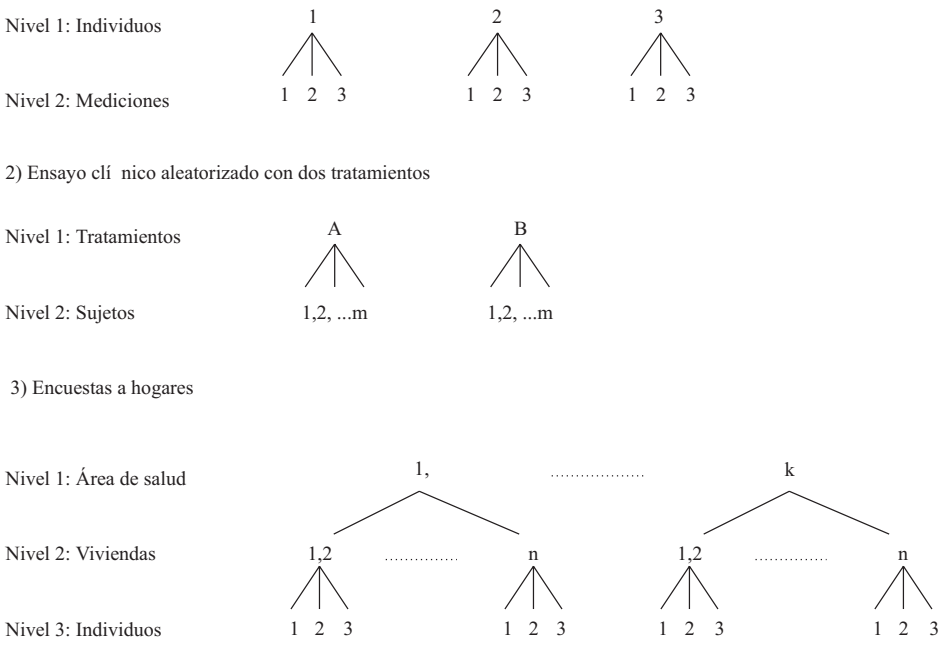
ma en que se vertebra la muestra. Ejemplos del primer tipo son los estudios de cohortes con medidas repetidas, donde se realizan varias mediciones a cada persona de la muestra durante el tiempo que dura el estudio. Otro ejemplo son los ensayos clínicos aleatorizados en los que grupos de individuos se someten a diferentes tratamientos. El segundo tipo se produce en las encuestas a hogares, en las que conglomerados geográficos de los hogares son muestreados para minimizar los costos de la investigación y facilitar la construcción del marco muestral (figura 1)[1].

El análisis de estos datos puede ser complicado por la existencia, al menos potencial, de correlación entre las observaciones de un mismo conglomerado. Esta correlación se debe tomar en cuenta para obtener estimaciones apropiadas de los parámetros, sus varianzas y la correcta distribución de los estadígrafos[2-4].

Durante mucho tiempo la práctica regular ha sido ignorar la estructura de los datos y aplicar los métodos estadísticos estándares para su análisis. Esto se ha debido, por una parte, a la escasa disponibilidad de programas informáticos capaces de calcular los estimadores apropiados mediante algoritmos complicados. Por otra parte, los paquetes estadísticos más usado como SAS, SPSS, STATISTICA, NCSS, etcétera, contemplan una gran diversidad de métodos estadísticos, incluidos los métodos multivariados,

Figura 1

Ejemplos de estudios con estructura de datos jerárquica o de conglomerados



pero las estimaciones de las varianzas son hechas obviando la estructura de los datos y suponiendo que provienen de un muestreo simple aleatorio.

A partir de los años 80 comienzan a aparecer en la literatura científica diferentes procedimientos que toman en cuenta la estructura de conglomerados en los datos[5-7]. Aunque los trabajos son dispersos y se enfocan para cada diseño particular, se pueden vislumbrar dos enfoques principales para el análisis de este tipo de datos: *análisis de promedios poblacionales* o *agregado,* en el que la heterogeneidad entre los conglomerados no se modela explícitamente, sino que se considera la correlación intra-conglomerado al calcular las estimaciones de los parámetros poblacionales. A los modelos bajo este enfoque se les conoce como *modelos marginales* porque calculan los estimadores de los parámetros de la regresión condicionados a la estructura de los datos. El otro enfoque es el *análisis de sujeto específico o desagregado,* en el que interesa modelar explícitamente la variabilidad entre los conglomerados y se calculan estimaciones para cada grupo[3,4].

Este trabajo es de naturaleza metodológica y se destina a ilustrar, en función de la precisión de las estimaciones, algunos métodos que pueden ser usados para estimar los parámetros poblacionales y los modelos de regresión en los datos con estructura jerárquica. Para ello se usaron datos de la *I Encuesta Nacional de Factores de Riesgo y Actividades Preventivas*[8], realizada en Cuba durante el año 1995. Además, se brindan recomendaciones generales para indicar situaciones en las cuales se cometen errores cuando se usan métodos y paquetes estadísticos estándares al analizar este tipo de datos.

MATERIAL Y MÉTODOS

Se emplearon los datos procedentes de la *I Encuesta Nacional de Factores de Riesgo y Actividades Preventivas*[8], realizada en Cuba durante el año 1995. La encuesta fue realizada con dos objetivos: describir en la población cubana los principales factores de riesgo asociados a las enfermedades crónicas no transmisibles y estudiar la correlación entre los factores de riesgo y las características socio-demográficas.

La encuesta se realizó en una muestra de 14.305 personas mayores de 15 años y residentes en el área urbana del país. Se utilizó un diseño muestral no equiprobabilístico en etapas que generó un conjunto de datos con estructura de conglomerados. Se consideraron como estratos los 169 municipios urbanos del país. Las unidades de la primera etapa de selección fueron los distritos y se seleccionaron con probabilidades proporcionales al número de casas particulares en los municipios. Se seleccionaron 740 distritos (entre 3 y 5 en cada estrato) y se visitaron 4.429 viviendas (entre 20 y 30 en cada estrato).

Los distritos se dividen en áreas de aproximadamente 64 viviendas; se seleccionó un área con probabilidad proporcional al número de casas particulares como unidad de segunda etapa. El área seleccionada se divide en secciones que tienen 4 casas como promedio y 2 de ellas se seleccionan con igual probabililidad de selección como unidades de tercera etapa. Dentro de cada sección seleccionada todas las casas son muestreadas y todos los residentes mayores de 15 años fueron entrevistados[8].

En esta encuesta se recogieron numerosas variables, pero para este artículo se decidió tomar 9 de ellas: condición de hipertenso, índice de masa corporal, sexo, edad, nivel educacional, tensión arterial diastólica y sistólica, hábito de fumar, actividad física y regiones. La descripción de la muestra se presenta en la tabla 1.

En este trabajo se abordan dos problemas. El primero está relacionado con la estimación de los parámetros poblacionales, sus errores estándares y las pruebas de significación en datos con estructura jerárquica. El segundo está dirigido a la estimación de los

Tabla 1

Medidas de resúmenes de las variables involucradas en la muestra

Variables	*Medidas de Resumen*
Condición de hipertenso	0,351* (0,0042)
Condición de fumadores	0,332* (0,0043)
Condición de sedentarios	0,334* (0,0041)
Sexo (Hombres)	0,480* (0,0044)
Índice de masa corporal: $\frac{Peso\ en\ kg}{(Talla\ in\ m)^2}$	23,60** (0,0450)
Edad en años	43,12** (0,1483)
Nivel educacional (pre-universitario o universitario)	0,381* (0,0040)
Tensión diastólica	78,48** (0,1016)
Tensión sistólica	124,13** (0,1656)

* Proporción de la categoría especificada en cada variable.
** Valor medio
Los valores entre paréntesis corresponden a los errores estándares

coeficientes en modelos de regresión para este tipo de datos.

Con relación a los parámetros poblacionales se ejemplificará con la estimación de proporciones y para las pruebas de significación se hará con la prueba χ^2. Los procedimientos a comparar serán los siguientes:

— **Análisis Convencional (AC):** asume que los datos provienen de una muestra simple aleatoria. Los estimadores que se usan son los convencionales. Para la prueba de independencia se utiliza el estadístico clásico χ^2 de Pearson.

— **Análisis Ponderado (AP):** se incorporan los pesos muestrales en el análisis para el cálculo de las estimaciones de los parámetros poblacionales, no así de los errores estándares. Este permite solucionar el problema de la diferencia entre las probabilidades de selección de los sujetos. Para la prueba de independencia se utiliza el estadístico clásico χ^2 de Pearson.

— **Análisis Ajustado (AA):** este procedimiento, además de considerar los pesos muestrales, toma en cuenta la estructura de conglomerados de los datos para el cálculo de las estimaciones de los parámetros poblacionales y sus errores estándares. En este caso la estimación de la proporción se obtiene por medio de los estimadores de razón y los errores estándares se calculan tomando en consideración la estructura de conglomerados del diseño[5,6]. La prueba χ^2 se basa en el estadígrafo de Wald y en la estrategia propuesta por Koch, Freeman y Freeman[9] para el análisis de muestras complejas.

El paquete estadístico SAS versión 7.0 se usó para aplicar los métodos AC y AP. Otros paquetes estándares (SPSS, STATISTICA) proveen resultados similares. SUDAAN versión 7.5 se empleó para aplicar el tercer método (AA).

Para ilustrar los métodos que pueden emplearse para estimar los coeficientes en los modelos de regresión se utilizaron los siguientes procedimientos:

— **Regresión Lineal Múltiple Ordinaria (RLMO):** el modelo de regresión usado es

$$y_i = \alpha + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \varepsilon_i \quad (1)$$

Para la estimación de los coeficientes de la regresión se asume que los datos provienen de una muestra simple aleatoria y se obtienen por el método de mínimos cuadrados ordinarios.

— **Regresión Lineal Múltiple con Pesos (RLMP):** el modelo de regresión es el mismo que el usado en el RLMO. La diferencia radica en la estimación de los coeficientes que se obtienen por el método de mínimos cuadrados ponderados, pero se sigue asumiendo que las observaciones son independientes.

— **Regresión Lineal Múltiple usando la Ecuación de Estimación Generalizada (GEE):** aquí el modelo sigue siendo el mismo que el de RLMO, pero para la estimación de los coeficientes y las pruebas de significación[8] se toma en cuenta la diferencia en las probabilidades de selección y la estructura de conglomerados de los datos.

— **Modelo de efectos aleatorios (MIXED):** un posible modelo puede ser el de interceptos aleatorios, donde los coeficientes de las variables explicativas son fijos y el término constante varía entre conglomerados. Este, además de considerar la estructura de conglomerados de los datos, permite modelar la variación entre ellos. En este caso el modelo usado es

$$\begin{cases} y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + ... + \beta_p x_{pij} + \varepsilon_{ij} & i = 1,...,I \\ \beta_{0j} = \beta_0 + \mu_j & j = 1,...,J \end{cases} \quad (2)$$

donde J es el número de conglomerados e I_j es el número de individuos dentro del conglomerado j-ésimo. Las estimaciones se realizan mediante el método de mínimos cuadrados ponderados[10].

Los tres primeros métodos siguen un enfoque agregado y el último está considerado como un método de análisis del enfoque desagregado. Se utilizó el paquete estadístico SAS para ejecutar los procedimientos RLMO, RLMP y MIXED, mientras que SUDAAN se empleó para aplicar el método GEE.

ESTIMACIÓN DE LOS PARÁMETROS POBLACIONALES

La estimación de la prevalencia de hipertensión arterial es sobrestimada en un 15% cuando se utilizan los estimadores convencionales comparados con los estimadores ponderados (obtenidos mediante AP y AA). Los errores estándares son menores mediante los procedimientos convencionales que los calculados cuando se toma en cuenta la estructura de conglomerados (para Cuba 0,3 vs. 0,6). Sin embargo, el análisis con pesos genera valores de los errores estándares mucho menores que los otros métodos. Esto se debe a que el SAS considera que el tamaño de la muestra es igual a la suma de los pesos, por ejemplo en Ciudad de La Habana la muestra sería de 1.734.434 personas en lugar de 2.069, que es el verdadero tamaño muestral (tabla 2).

Tabla 2

Prevalencia de hipertensión arterial y sus errores estándares por regiones y procedimientos de análisis

Región	*Análisis Convencional*	*Análisis con Pesos*	*Análisis Ajustado*
Ciudad de La Habana	34,7 (0,9)	30,2 (0,03)	30,2 (1,3)
Occidente	36,6 (0,7)	31,9 (0,04)	31,9 (1,1)
Centro	38,5 (0,7)	33,5 (0,05)	33,5 (1,4)
Oriente	31,5 (0,6)	27,4 (0,03)	27,4 (0,9)
Cuba	*35,1 (0,3)*	*30,5 (0,04)*	*30,5 (0,6)*

En la tabla 3 se muestran los resultados al aplicar la prueba χ^2 para contrastar la hipótesis de que no existe asociación entre la hipertensión y el sexo. El valor del estadígrafo, para Cuba, obtenido mediante el análisis convencional es mayor que el calculado mediante el análisis ajustado. Esto implica que la probabilidad asociada sea menor (0,004 contra 0,01). Dentro de las regiones no se encontraron patrones, en algunos casos son mayores y en otros menores. Al realizar el análisis con pesos los valores de los estadígrafos son mucho mayores en Cuba (4.757,9 vs 6,6 en AA y 8,3 en AC) y en las regiones, lo cual se debe al tamaño inflado de la muestra usado para su cálculo.

ESTIMACIÓN DE LOS COEFICIENTES EN LOS MODELOS DE REGRESIÓN

Las estimaciones de los coeficientes para el modelo (1) obtenidos a través de los métodos RLMO, RLMP y GEE se muestran en

Tabla 3

Estadígrafo de la prueba χ^2 y el valor p asociado para la evaluación de la asociación entre hipertensión y sexo

Región	*Análisis Convencional*	*Análisis con Pesos*	*Análisis Ajustado*
Ciudad de La Habana	0,3 (0,6)	128,7 (0,001)	0,2 (0,7)
Occidente	0,5 (0,5)	373,8 (0,001)	1,1 (0,3)
Centro	2,9 (0,09)	2.260,1 (0,001)	3,1 (0,08)
Oriente	8,3 (0,004)	3.586,9 (0,001)	4,2 (0,04)
Cuba	*8,3 (0,004)*	*4.757,9 (0,001)*	*6,6 (0,01)*

la tabla 4. Al comparar estos tres métodos se aprecia que la estimación de los coeficientes y los errores estándares obtenida por cada método es diferente. Se aprecia que el método RLMP es el que produce los errores estándares más bajos y GEE los más altos.

Con RLMO y RLMP las variables que resultaron significativas son sexo, nivel de escolaridad, tensión arterial diastólica y tensión arterial sistólica, hábito de fumar y ser sedentario o no. Con GEE se encontró que la edad, el sedentarismo y el nivel educacional no tienen un efecto significativo en el índice de masa corporal.

Los resultados de aplicar un modelo en dos etapas (MIXED) se muestran en la ta-

Tabla 4

Estimaciones de los coeficientes de la regresión del IMC y sus errores estándares en los modelos marginales

Variables	*RLMO*	*RLMP*	*GEE*
Edad en años	0,0019	0,0050	0,0016
	0,0036	0,0033	0,0049
	0,6044	0,1311	0,7429
Hombres	–0,8399	–0,8696	–0,8681
	0,0999	0,0914	0,1153
	0,0001	*0,0001*	*<0,0001*
Nivel educacional	0,1061	0,0923	0,0708
	0,0286	0,0254	0,0412
	0,0020	*0,0003*	*0,0861*
Sedentarios	–0,3277	–0,2352	–0,2380
	0,1024	0,0967	0,1299
	0,0014	*0,0151*	*0,0675*
Fumadores	–0,8559	–0,8841	–0,8138
	0,1012	0,0926	0,1347
	0,001	*0,0001*	*<0,0001*
Tensión diastólica	0,0754	0,0553	0,0581
	0,0062	0,0058	0,0109
	0,0001	*0,0001*	*<0,0001*
Tensión sistólica	0,0282	0,0337	0,0331
	0,0040	0,0037	0,067
	0,0001	*0,0001*	*<0,0001*

Cada celda contiene el coeficiente de la regresión, el error estándar y los valores de p, respectivamente.

bla 5. Se consideró como unidad de segundo nivel los municipios donde residen los encuestados. El 91,3% de la variabilidad total es explicada por las variables individuales y el 8,7% es atribuible a las unidades superiores (municipios) y es altamente significativa. Aquí se aprecia que todas las variables excepto la edad y sexo resultaron significativas.

Tabla 5

Estimaciones de los coeficientes de la regresión del IMC y sus errores estándares en el modelo de interceptos aleatorios

Variables	*Coeficientes*	*Error estándar*	*Valores p*
Fijos:			
Edad (en años)	–0,0014	0,0033	0,6676
Hombres	–0,7638	0,0868	0,1440
Nivel educacional	0,0775	0,0263	0,0032
Sedentarios	–0,3234	0,0952	0,0007
Fumadores	–0,8545	0,0914	<0,0001
Tensión diastólica	0,0711	0,0057	<0,0001
Tensión sistólica	0,0293	0,0037	<0,0001
Aleatorios:	4,6425	0,5460	<0,0001

COMENTARIOS

Aunque la evidencia empírica en nuestro trabajo está basada sólo en un ejemplo, los hallazgos son consistentes con los aportados por otros estudios[11-13]. Cuando se usa información procedente de diseños muestrales complejos es necesario prestar una atención especial a la selección de los estimadores apropiados para los parámetros objeto de inferencias.

Estimación de los parámetros poblacionales

No existen muchas dificultades para obtener estadísticos ponderados puesto que los paquetes estadísticos estándares contemplan esta opción. Sin embargo, estos ofrecen estimaciones no sesgadas de los parámetros pero no de sus errores estándares. De los tres procedimientos aplicados para la estimación de los parámetros poblacionales en el ejemplo, los que usan los pesos (AP y AA) proporcionan una estimación no sesgada de ellos. Esto se debe a que no todos los individuos tienen iguales probabilidades de pertenecer a la muestra y los pesos muestrales incorporan estas diferencias, así como las de cobertura del muestreo y debidas a la no respuesta en algunos aspectos. La magnitud del sesgo de los estimadores convencionales depende de los datos y de los métodos empleados para el cálculo de los pesos muestrales[2,14].

En el AC, además del sesgo que induce en las estimaciones puntuales, los errores estándares y las medidas de variabilidad generalmente son subestimadas. Esto se debe a que no toma en cuenta la estructura de conglomerado de los datos y las diferencias en las probabilidades de selección. Las estimaciones de los errores estándares más eficientes son las que se obtienen por el AA, esto se debe a que este método toma en cuenta la estructura de conglomerados para calcular las estimaciones de las varianzas. Sin embargo, en el análisis con pesos se obtienen las estimaciones de las varianzas más sesgadas, por el tamaño de muestra artificial que se emplea en su cálculo. Algo similar ocurre al aplicar las pruebas de hipótesis debido a que las varianzas son subestimadas, lo que provoca que se obtengan diferencias estadísticamente significativas cuando en realidad no las hay.

Estimación de los coeficientes en los modelos de regresión

En las investigaciones sanitarias son frecuentes los estudios en los que no se cumple la hipótesis de independencia entre los valores de la variable resultado, necesaria para poder utilizar los modelos de regresión convencionales. Ejemplos de estos estudios son los transversales con muestras complejas,

estudios de cohorte con medidas repetidas y los ensayos clínicos cruzados (cross-over), entre otros.

Cabe señalar que el mecanismo de selección de la muestra no sólo es importante cuando se quieren realizar inferencias poblacionales, sino para identificar relaciones o asociaciones entre variables por medio de modelos de regresión.

Bajo un enfoque agregado el modelo GEE es el que genera las estimaciones de los coeficientes y sus errores estándares más fiables, debido a que este método toma en cuenta las diferencias en las probabilidades de selección y la estructura de conglomerado o jerárquica que se genera en los datos producto del diseño. A la vez el modelo menos fiable es el RLMO pues no toma en consideración ninguno de estos dos aspectos. El modelo RLMP a pesar de que no considera la estructura de conglomerados utiliza para la estimación de los coeficientes los pesos muestrales.

La suposición de que los sujetos que pertenecen a distintos conglomerados tienen características diferentes es considerado tanto bajo un enfoque agregado como en el desagregado, pero la manera de abordarlo es diferente. En el primero se trata de «promediar» estas diferencias y en el segundo se modelan explícitamente, al plantear el problema en niveles. Este último enfoque es más flexible, pues aquí no sólo interesa la eficiencia estadística sino la lógica de afrontar el problema de investigación, pues descomponer la varianza en los distintos niveles permite identificar qué parte de la variabilidad total corresponde a cada uno de ellos. En nuestro ejemplo se encontraron diferencias atribuibles a residir en uno u otro municipio del país y estas diferencias son estadísticamente significativas. Cabe señalar que éstas no pueden ser identificadas en un modelo marginal debido a que en este enfoque no se modelan explícitamente las variaciones entre los conglomerados.

Con el modelo de interceptos aleatorios se encontró que los coeficientes de las variables individuales son constantes entre los conglomerados, pero la variabilidad entre los conglomerados sugiere que características de los municipios tienen un efecto en los niveles de la variable de respuesta (índice de masa corporal). Este es el modelo más sencillo, pues se pueden ajustar otros modelos como uno de coeficientes aleatorios, lo que permite que los coeficientes varíen entre los conglomerados. O bien un modelo más general donde además se añaden variables en los niveles superiores[15,16].

Aunque el modelo GEE genera estimaciones fiables no permite identificar la existencia de variabilidad en niveles superiores. Con los modelos de efectos aleatorios no sólo se modela el efecto de un conjunto de variables individuales en la variable de respuesta, sino que este modelo permite identificar qué proporción de la varianza total es atribuible a los niveles superiores.

Los métodos de análisis que contemplan la estructura de los datos y las características del diseño dan una versión más realista del problema bajo estudio. Con ellos se obtienen estimaciones más precisas de los parámetros y sus errores estándares. Al ajustar modelos de regresión no sólo importa obtener eficiencia en la estimación de los coeficientes sino que debe considerarse el enfoque (agregado o desagregado) para modelar el problema de investigación. La frecuencia con la que se encuentran este tipo de datos en los estudios epidemiológicos y de salud pública demandan una mayor utilización de estos métodos y de los paquetes estadísticos que los contemplan.

BIBLIOGRAFÍA

1. Graubard BI & Korn EL. Regression analysis with correlated data. Statistics in Medicine 1994; 13: 509-522.

2. Brogan DJ. Pitfalls of using standard statitistical *software packages for sample survey data.* En:

Peter Armitage, editor. Encyclopedia of Biostatistics. John Wiley & Sons, New York; 1998.

3. Diggle PJ, Liang KY y Zeger SL. Analysis of Longitudinal Data. Oxford University Press; 1994.

4. Skinner CJ, Holt D y Smith TMF. Analysis of complex surveys. New York: John Wiley; 1989.

5. Hidiriglou MA y Rao JNK. Chi-Squared tests with categorical data from Complex Surveys: Part I. Journal of Official Statistics 1987; 3(2): 117-132.

6. Hidiriglou MA y Rao JNK. Chi-Squared tests with categorical data from Complex Surveys: Part II. Journal of Official Statistics 1987; 3(2): 133-140.

7. Binder DA. On the variances of asymptotically normal estimators from complex surveys. International Statistics Review 1983; 51: 279-292.

8. Bonet M *et al.* Encuesta Nacional de Factores de Riesgo y Conductas Preventivas, Cuba. Reporte de Investigación. La Habana: Instituto Nacional de Higiene, Epidemiología y Microbiología; 1995.

9. Koch GG, Freeman DH Jr y Freeman JL. Strategies in the multivariate analysis of data from complex surveys. International Statistics Review 1975; 43: 59-78.

10. Verbeke G y Molenberghs G. Linear mixed models practice. New York: Springer-Verlag; 1997.

11. Donner A y Bull S. Inferences concerning a common intraclasss correlation coefficient. Biometrics 1983; 39: 771-775.

12. Donner A y Donald A. Analysis of data arising from a stratified design with the cluster as unit of randomization. Statistics in Medicine 1987; 6: 43-52.

13. Mian IUH y Shoukri MM. Statistical analysis of intraclass correlations from multiple samples with applications to arterial blood pressure data. Statistics in Medicine 1997; 16: 1497-1514.

14. Warszawski J, Messiah A, Lellouch J, Meyer L y Deville JC. Estimating means and percentages in a complex sampling survey: application to a french national survey on sexual behaviour (ACSF). Statistics in Medicine 1997; 16: 397-423.

15. Albert PS. Tutorial in Biostatistics. Longitudinal data in analysis (repeated measures) in clinical trials. Statistics in Medicine 1999; 18: 855-88.

16. Sullivan LM, Dukes KA y Losina E. Tutorial of Biostatistics. An introduction to hierarchical linear modelling. Statistics in Medicine 1997; 18: 397-423.