Efecto del diseño muestral en el análisis de encuestas de diseño complejo. Aplicación a la encuesta de salud de Catalunya

M. Guillén[1] / S. Juncà[2] / M. Rué[3] / J. M. Aragay[1]
[1]Departament d'Econometria, Estadística i Economia Espanyola, Facultat de Ciències Econòmiques, Universitat de Barcelona.
[2]Servei Català de la Salut.
[3]Servei d'Epidemiologia Clínica i Salut Pública, Hospital de Sant Pau, Barcelona.

Correspondencia: Montserrat Guillén. Dept. d'Econometria, Estadística i Economia Espanyola. Facultat de Ciències Econòmiques i Empresarials. Universitat de Barcelona. Tte. Cor. Valenzuela, 1-11. 08034 Barcelona. E-mail: guillen@eco.ub.es

(Effect of sample design on the analysis of complex surveys. Application to the Catalonia health survey, Spain)

Resumen

Objetivo: Ilustrar cómo se puede incorporar fácilmente el diseño muestral en el análisis estadístico de encuestas con diseños muestrales complejos, para obtener estimaciones correctas.

Métodos: Se ha utilizado un programa estadístico (STATA) que permite incorporar el diseño muestral para analizar la encuesta de salud de Cataluña del año 1994.

Resultados: Si se tiene en cuenta el diseño muestral completo, las estimaciones son insesgadas. Si únicamente se tienen en cuenta las ponderaciones individuales se obtienen estimaciones puntuales insesgadas pero se suelen subestimar los errores estándar.

Conclusiones: La disponibilidad actual de programas estadísticos que permiten incorporar el diseño muestral de manera sencilla debería animar a los investigadores a analizar mejor las encuestas.

Summary

Objective: To illustrate the way in which a sample design can be easily incorporated into the statistical analysis of complex surveys to obtain correct estimates.

Methods: A statistical program (STATA) was used to analyze the Catalan Health Survey (Spain) for the year 1994.

Results: If the sample design is taken into account, the estimates are unbiased. If only the weights are taken into account, unbiased point estimates are obtained, but the standard errors tend to be underestimated.

Conclusions: The current availability of statistical programs that allow to incorporate easily the sample design should encourage researchers to better analyze their surveys.

Introducción

Los estudios epidemiológicos a menudo utilizan datos obtenidos mediante encuestas. A fin de optimizar el proceso de obtención de los datos, con frecuencia, las muestras seleccionadas no son muestras aleatorias simples, sino muestras de diseño complejo que tienen una o varias de las características siguientes[1]:

Estratos. Son una partición de la población de estudio. En cada uno de los estratos se seleccionan las unidades muestrales (individuos o grupos) de manera independiente.

Conglomerados: Previamente a la selección de los individuos muchas encuestas seleccionan grupos aleatoriamente (p. ej. municipios, escuelas, hospitales, etc.). A diferencia de los estratos, no se seleccionan individuos en todos los conglomerados, sino solamente en los conglomerados que han sido seleccionados en etapas previas. Puede haber una o varias etapas de muestreo antes de llegar a la selección final de los individuos.

Las respuestas de los individuos de un mismo grupo (estrato o conglomerado) suelen ser más parecidas entre sí que las de individuos de otros grupos.

Ponderaciones: Una poderación w_j para la observación *i* significa que la observación *i* representa a w_j elementos en la población. Es muy importante calcular correctamente las ponderaciones, ya que éstas son una consecuencia del método de muestreo. Equivalen, para cada individuo, al inverso de la probabilidad de ser seleccionado.

Por ejemplo, una ciudad se puede compartimentar en distritos (estratos), en cada distrito seleccionar un grupo de áreas básicas de salud (conglomerados) y en cada área básica de salud una muestra aleatoria de pacientes.

Hasta hace poco, era difícil analizar datos de encuestas de diseño complejo con los programas esta-

dísticos más utilizados. Estos programas suelen asumir que los datos proceden de una muestra simple aleatoria y son independientes e idénticamente distribuidos. Su utilización en muestras de diseño complejo produce resultados sesgados.

Varios autores han estudiado las técnicas que deben emplearse para obtener estimaciones insesgadas en muestras de diseño complejo[2,3]. Utilizando técnicas de remuestreo, es posible abordar el cálculo de los errores estándar en situaciones en las que la complejidad del diseño no permiten una aproximación sencilla[4]. Actualmente, programas como SUDAAN[5], PC CARP[6], y STATA[7] entre otros, utilizan aproximaciones lineales que permiten estimar fácilmente los parámetros poblacionales y sus errores estándar, incorporando el diseño muestral.

Este trabajo pretende ilustrar, tomando como ejemplo la Encuesta de Salud de Catalunya (ESCA) del año 1994, cómo se puede incorporar fácilmente el diseño muestral en el análisis estadístico para obtener estimaciones insesgadas. Asimismo, compara los resultados obtenidos al incorporar el diseño muestral con los que se hubieran obtenido: 1) teniendo en cuenta únicamente las ponderaciones de los individuos, y 2) considerando que la muestra es aleatoria simple, sin ponderar.

Material y métodos

La Encuesta de Salud de Catalunya (ESCA) del año 1994 tenía como objetivo proporcionar información fiable para cada una de las ocho regiones sanitarias que integran el Servei Català de la Salut y, por agregación, de la población no institucionalizada residente en la Comunidad Autónoma de Catalunya[8].

En el proceso de diseño se tomó en consideración los costes derivados de la obtención de la información, realizada por entrevista personal en el domicilio particular, y la disparidad en el tamaño poblacional de las ocho regiones, que implicaba cierta dificultad en hallar una única fracción de muestreo que fuera satisfactoria para todas ellas. El diseño de la ESCA fue bietápico: en la primera etapa, se realizó una estratificación por regiones sanitarias con fracciones de muestreo distintas y, dentro de cada una de ellas, se efectuó una selección según el tamaño poblacional de los municipios. Las unidades muestrales seleccionadas en la primera etapa fueron los municipios (conglomerados). En la segunda etapa, se eligieron los individuos a entrevistar en cada municipio, obteniéndose la información directamente de la persona escogida o de un adulto en el caso de los niños o incapacitados para responder. Se entrevistaron en total 15.000 individuos.

El trabajo realiza un recorrido, mediante ejemplos, por diversos procedimientos estadísticos. En cada situación se exponen los resultados obtenidos mediante la incorporación de todas las características relevantes del diseño muestral. Concretamente, se ha tenido en cuenta la estratificación de la muestra en ocho regiones, la selección de municipios en primer lugar e individuos en segundo y la inclusión de las ponderaciones individuales. A continuación, dado que la mayoría de programas estadísticos disponen de la opción de ponderación, aunque no sea posible tener en cuenta otras características de la muestra, se presentan los resultados obtenidos incorporando únicamente las ponderaciones e ignorando el resto de características del diseño. Finalmente, se presentan los resultados obtenidos bajo la hipótesis de que los datos provienen de una muestra aleatoria simple, sin ponderar.

Los resultados que se presentan se han obtenido mediante el programa estadístico STATA. Para estimar la varianza en los diseños muestrales complejos, el programa STATA, al igual que otros programas como SUDAAN y PC CARP, construye una aproximación lineal del estadístico de interés, y calcula el error estándar de dicha aproximación[9]. Los comandos de STATA que permiten incorporar el diseño muestral son:

— *svyset strata* región
— *svyset psu* municipio
— *svyset pweight* peso

Para obtener proporciones, medias, regresión lineal y regresión logística, se pueden utilizar los comandos *svytab*, *svymean*, *svyreg* y *svylogit*. Los programas utilizados en los ejemplos que se presentan se pueden consultar en la página web *http://www.eco.ub.es/~guillen/esca*.

Resultados

Estimación

El porcentaje (intervalo de confianza al 95%) de personas que declaran tomarse la tensión periódicamente, de forma preventiva, en la región sanitaria de Tarragona es 29,9 (25,0-34,9), si se tiene en cuenta el diseño muestral completo; 29,9 (27,5-32,4) si solamente se tienen en cuenta las ponderaciones, y 30,2 (27,7-32,6) si se supone que la muestra es aleatoria simple. Mientras que las estimaciones puntuales del porcentaje son iguales o muy similares en los tres tipos de muestreo, la precisión de la estimación depende del método utilizado. Al tener en cuenta el diseño muestral, el intervalo de confianza del porcentaje es más amplio, hecho que refleja que el error estándar estaba subestimado a causa de la asociación entre los individuos que pertenecen al mismo municipio.

Contraste de hipótesis: comparación de medias y comparación de proporciones

Se han comparado las medias de las puntuaciones de la escala visual analógica del instrumento de medida de la calidad de vida EUROQOL (escala de 0 a 100) y la prevalencia de diabetes según el género del entrevistado (tabla 1). Los intervalos de confianza de las medias son más amplios si se tiene en cuenta el diseño muestral. Sin embargo, en este ejemplo, la prueba de hipótesis da el mismo resultado en los tres escenarios muestrales. En cambio, al comparar el porcentaje de personas que declaran ser diabéticos, según el género, se obtienen resultados diferentes en el contraste de hipótesis según se tenga en cuenta o no el diseño muestral. Nótese que cuando el muestreo no es aleatorio simple, el estadístico de contraste para variables categóricas no sigue una distribución ji-cuadrado, sino F de Snedecor.

Modelos de regresión

En la tabla 2 se presentan los resultados obtenidos al utilizar modelos de regresión lineal y logística. En el modelo de regresión lineal se intenta explicar el valor de la escala visual analógica del instrumento EUROQOL en función de la edad del individuo. A fin de poder apreciar un efecto desacelerador, se incluye también la edad al cuadrado. El modelo se ha ajustado para el colectivo de hombres de la encuesta. En la tabla se presentan los coeficientes, errores estándar y valores p en las tres situaciones analizadas. Para ilustrar los resultados del modelo de regresión logística se presentan las *odds ratio* y sus intervalos de confianza, para la variable dependiente que indica si el entrevistado declara realizarse un examen médico sistemático de forma preventiva. Se toman como variables explicativas la edad, el género y la clase social.

En todos los casos se observan diferencias en las estimaciones puntuales entre el método que incorpora el diseño completo con el método que asume muestreo aleatorio simple. Los intervalos de confianza son más amplios al incorporar el diseño muestral completo. Por este motivo, algunas variables que eran estadísticamente significativas dejan de serlo al tener en cuenta el diseño muestral.

Conclusiones

Los resultados muestran que, en todas las ocasiones, ignorar el diseño muestral conduce a estimaciones sesgadas de los parámetros de interés. Si únicamente se tienen en cuenta las ponderaciones individuales, las estimaciones puntuales son insesgadas pero los errores estándar, en general, están subestimados.

Dado que los programas estadísticos utilizados habitualmente incorporan la posibilidad de incluir ponderaciones, realizar esta operación resulta sencillo. No obstante, la disponibilidad actual de programas estadísticos que permiten incorporar el diseño muestral completo de manera sencilla, debería animar a los investigadores a analizar mejor las encuestas.

Tabla 1. Comparación de las medias de la escala visual analógica del instrumento EUROQOL (calidad de vida) y de la prevalencia de diabetes, según el género del entrevistado.

Género	Análisis que incorpora el diseño completo			Análisis que incorpora únicamente las ponderaciones			Análisis que asume muestreo aleatorio simple		
	EUROQOL								
	Media	E.E.	IC 95%	Media	E.E.	IC 95%	Media	E.E.	IC 95%
Hombres	75,5	0,34	74,8-76,2	75,5	0,23	75,1-76,0	75,3	0,21	74,9-75,7
Mujeres	71,3	0,40	70,5-72,1	71,3	0,25	70,8-71,8	71,2	0,22	70,8-71,7
Test	t = 12,5, p < 0,001			t = 12,5, p < 0,001			t = 13,5, p < 0,001		
Género	Prevalencia de diabetes								
	(%)	E.E. (%)	IC 95%	(%)	E.E. (%)	IC 95%	(%)	E.E. (%)	IC 95%
Hombres	4,4	0,28	3,9-5,0	4,4	0,31	3,8-5,0	4,5	0,27	4,1-5,0
Mujeres	5,1	0,32	4,4-5,7	5,1	0,30	4,5-5,7	5,7	0,28	5,1-6,2
Test	F = 1,88 p = 0,17			F = 2,08 p = 0,15			Ji-cuadrado = 8,8 p = 0,003		

E.E.: error estándar. OR: odds ratio. IC 95%: Intervalo de confianza al 95%.

Tabla 2. Comparación de resultados de los modelos de regresión lineal y logística según se tenga en cuenta o no el diseño muestral.

Modelo de regresión lineal. Variable dependiente: puntuación en la escala visual analógica EUROQOL. Hombres.

	Análisis que incorpora el diseño completo			Análisis que incorpora únicamente las ponderaciones			Análisis que asume muestreo aleatorio simple		
	β	E. E.	p	β	E. E.	p	β	E. E.	p
Edad (años)	–0,263	0,043	< 0,001	–0,263	0,036	< 0,001	–0,256	0,032	< 0,001
Edad[2]	–0,00083	0,00051	0,109	–0,00083	0,00045	0,066	–0,00097	0,00038	0,012
Constante	86,91	0,68	< 0,001	86,91	0,58	< 0,001	86,88	0,56	< 0,001
		R^2 = 0,191			R^2 = 0,191			R^2 = 0,188	

Modelo de regresión logística. Variable dependiente: realizar un examen médico preventivo

	Análisis que incorpora el diseño completo		Análisis que incorpora únicamente las ponderaciones		Análisis que asume muestreo aleatorio simple	
	OR	IC 95%	OR	IC 95%	OR	IC 95%
Edad[A]	1,008	1,005-1,011	1,008	1,005-1,010	1,008	1,006-1,010
Género[b]	1,066	0,925-1,227	1,066	0,964-1,178	1,115	1,022-1,215
Clase social[c]						
II	0,636	0,473-0,855	0,636	0,496-0,815	0,595	0,480-0,737
III	0,832	0,605-1,143	0,832	0,654-1,058	0,853	0,691-1,053
IVa	0,522	0,376-0,725	0,522	0,415-0,658	0,530	0,433-0,649
IVb	0,476	0,343-0,662	0,476	0,372-0,611	0,475	0,383-0,590
V	0,444	0,307-0,644	0,444	0,338-0,584	0,479	0,377-0,608
Otras	0,460	0,262-0,809	0,460	0,306-0,693	0,468	0,328-0,668

β: coeficientes del modelo de regresión. E. E.: error estándar. OR: odds ratio IC 95%: Intervalo de confianza al 95%. [a]: OR para incrementos de un año de edad. [b]: Categoría de referencia, hombres. [c]: Categoría de referencia, clase social I.

Agradecimientos

Los autores agradecen al Dr. Jorge Morel las sugerencias y apoyo que les brindó durante la exposición de parte de este trabajo en el *XVI Meeting of the International Society for Clinical Biostatistics*, celebrado en Barcelona del 31 de julio al 4 de agosto de 1995. A Mònica Bécue, Lluís de Jover, Esteve Fernández y Anna Schiaffino por sus comentarios.

Bibliografía

1. Skinner CJ, Holt D, Smith TMF, eds. Analysis of complex surveys. New York: Wiley; 1989.
2 Cochran WG. Sampling Techniques, 3 ed.. New York: John Wiley & Sons; 1977.
3. Wolter KM. Introduction to variance estimation. New York: Springer; 1985.
4. Murillo C, Guillén M. Estimación de las varianzas de las variables de la encuesta de salud de Barcelona. Gac Sanit 1989; 12:409-19.
5. Shah BV, Barnwell BG, Bieler GS, eds. SUDAAN user's manual, release 7.0. Research Triangle Park, NC: Research Triangle Institute; 1996.
6. Fuller WA, Kennedy W, Schnell D, et al. PC CARP. Ames, IA: Statistical Laboratory, Iowa State University; 1996.
7. Stata Corporation. Stata statistical software: release 5.0. College Station, TX: Stata Corporation; 1997.
8. Servei Català de la Salut. Àrea Sanitària. Enquesta de salut de Catalunya 1994. Generalitat de Catalunya; 1996.
9. Eltinge JL, Sribney WM. Stata Technical Bulletin 1996;31:3-42.