Isern, Guillerminia
Cuesta, Cristina
Barbona, Ivana
Meroi, Norma
Instituto de Investigaciones Teóricas y Aplicadas en Estadística (IITAE), Escuela de Estadística

COMPARACIÓN DE MODELOS PARA EL CONTROL DE SESGOS EN ESTUDIOS ECOLÓGICOS[1]

Resumen:

Entre los tipos de estudios más frecuentemente utilizados en epidemiología se encuentran los estudios ecológicos, cuya unidad de análisis es un área geográfica o un momento del tiempo y su objetivo es medir la asociación entre una enfermedad y un posible factor de riesgo o variable explicativa en base a medidas resumen observadas en esas unidades (promedio, porcentaje, tasa, etc.). Estos estudios son muy encontrados en la literatura debido a su sencillez, bajo costo, fácil obtención de la información, etc. y son generalmente mostrados en una faz exploratoria. El mayor problema que ellos presentan es la denominada "falacia ecológica", el sesgo que se comete al querer extrapolar las asociaciones observadas a nivel de área, a los individuos de las mismas. Este tipo de sesgos puede producirse debido a diferentes motivos: la incapacidad de determinar la temporalidad de la variable respuesta y la explicativa, la posible presencia de variables confundentes, la imposibilidad de contar con las distribuciones de las medidas agrupadas, etc. A fin de controlar el posible sesgo asociado a estos estudios se han propuesto diferentes modelos estadísticos, entre ellos los que tienen en cuenta estratos definidos de acuerdo a alguna posible variable de confusión. La ventaja de estos modelos estratificados es que se trabaja con grupos más pequeños que el total del área geográfica permitiendo tener en cuenta posibles variables de confusión (que son las que determinan los estratos). Sin embargo, esta metodología puede no ser adecuada cuando se carece de la información a nivel de estrato y/o cuando las áreas de estudio son demasiado pequeñas. Otra opción consiste en ajustar un modelo que pretenda controlar el mencionado sesgo combinando distintas fuentes de datos, una que incluye los datos agrupados y otra que incluye los datos a nivel individual (que puede no ser la misma que la anterior). Este modelo es un intermedio entre utilizar datos a nivel individuo y datos a nivel área geográfica. Se ejemplifica la metodología utilizando datos de la Encuesta Nacional de Factores de Riesgo llevada a cabo por las Direcciones Provinciales de Estadística entre octubre y diciembre de 2013. Se estudia la asociación entre obesidad y diferentes factores tales como edad, sexo, nivel de instrucción, cobertura de salud, etc. Se comparan los resultados de modelos para datos individuales con modelos para datos agrupados a nivel provincia o datos agrupados a nivel de estrato. Se concluye que los datos agrupados no logran reflejar las asociaciones

[1] Este trabajo se elaboró en el marco del Proyecto ECO164 Titulado "Modelos de ajuste y predicción para datos espacio-temporales", dirigido por Cristina Cuesta

observadas a nivel individual. A partir de este trabajo se dirige la investigación hacia el estudio de modelos que combinan tanto fuente de información a nivel individual como a nivel agrupado simultáneamente.

Palabras claves: falacia ecológica, datos agrupados, sesgos ecológicos

Introducción

Los estudios ecológicos se diferencian de otros fundamentalmente por su unidad de análisis y observación. Comúnmente éstas son áreas geográficas para las cuales se comparan las tasas o frecuencias de una enfermedad. La principal ventaja de estos estudios radica en la facilidad para disponer de datos agrupados o medidas resumen de las unidades a bajo costo y en poco tiempo. Sin embargo, una de las mayores limitaciones es el sesgo que puede ocurrir debido a que una asociación observada entre variables en los grupos no necesariamente representa la asociación que existe en el ámbito individual. Esto se conoce como “falacia ecológica”. Por su estructura, estos diseños no pueden extender sus hallazgos al caso individual. Mayormente el sesgo ocurre debido a la dificultad de controlar factores de confusión. Una estrategia sugerida para minimizar este sesgo es ajustar modelos teniendo en cuenta estratos (por ejemplo estratos construidos a partir de edad y sexo). La ventaja de la estratificación radica en trabajar con grupos más pequeños que el total del área geográfica lo cual aporta mayor información al análisis y conduce a resultados más similares a los que se obtendrían trabajando a nivel individual, pero también puede surgir el problema de la inestabilidad en áreas pequeñas. En este trabajo se comparan distintas estrategias de análisis utilizando datos de la Encuesta Nacional de Factores de Riesgo realizada en el año 2013 por el INDEC y el Ministerio de Salud de la Nación. Se modela la asociación entre obesidad y diferentes factores de riesgo recolectados en la encuesta, características sociodemográficas, educativas, etc. Las alternativas de análisis se refieren al ajuste de los modelos utilizando:

- datos a nivel individual (incluyendo entre las variables explicativas a la provincia, como efecto fijo y como efecto aleatorio),
- como unidad de análisis a la provincia y como variables explicativas, las mismas que se consideraban a nivel individual pero a nivel agrupado (porcentajes para variables categóricas y promedios para variables continuas),
- como unidad de análisis los estratos formados por las combinaciones de grupos de edad y sexo en cada provincia y como variables explicativas, las mismas que se consideraban a nivel agrupado,

El relevamiento de la encuesta fue realizado entre los meses de octubre y diciembre de 2013 por cada una de las Direcciones Provinciales de Estadística. La población objetivo eran personas de 18 años o más que habitan hogares particulares en localidades de 5.000 o más habitantes. Se obtuvo una muestra de 46.555 viviendas a nivel país. Todas las provincias (incluso Ciudad Autónoma de Buenos Aires) quedaron representadas. La tasa de respuesta fue del 70.7%.

De las variables relevadas en la encuesta, se utilizaron para este trabajo: obesidad (si,no;

basada en el Índice de Masa Corporal, si el IMC es mayor o igual a 30 se considera obesidad), sexo, edad, nivel de instrucción (hasta primario incompleto, primario completo o secundario incompleto, secundario completo y más) , cobertura de salud (con obra social o prepaga, sin obra social o prepaga), nivel de actividad física (intenso, moderado, bajo), consumo de tabaco (fumador, ex fumador, no fumador), prevalencia de hipertensión arterial (si, no), consumo de al menos 5 porciones diarias de frutas y/o verduras en una semana típica (si, no), prevalencia de hipercolesterolemia (si, no) y prevalencia de diabetes (si, no).

Metodología

Los modelos que se presentan continuación son los que se utilizaron para estudiar la relación entre obesidad y distintos factores de riesgo. Se presentan tres grupos de modelos, los que hacen uso de la información a nivel individual, a nivel agrupado (por provincia y/o estrato) o ambas agrupaciones.

- **Modelos para datos a nivel individual**

1) Modelo Logístico, sin efecto asociado a las provincias

$$p_i = \frac{\exp\left(\beta_0 + \sum_j x_{ij}\beta_j\right)}{1+\exp\left(\beta_0 + \sum_j x_{ij}\beta_j\right)} \qquad \text{(Modelo 1)}$$

donde p_i es la probabilidad del evento del individuo "i", x_{ij} es el valor de la covariable "j" asociada al individuo "i"

2) Modelo Logístico, con efecto fijo asociado a las provincias

$$p_{ik} = \frac{\exp\left(\beta_0 + \sum_j x_{ijk}\beta_j + \gamma_k\right)}{1+\exp\left(\beta_0 + \sum_j x_{ijk}\beta_j + \gamma_k\right)} \qquad \text{(Modelo 2)}$$

p_{ik} es la probabilidad del evento del individuo "i" en la provincia "k", x_{ijk} es el valor de la covariable "j" asociada al individuo "i" en la provincia "k", γ_k es el efecto fijo asociado a la provincia "k"

3) Modelo Logístico, con intercepto aleatorio asociado a las provincias

$$p_{ik} = \frac{\exp\left(\beta_{0k} + \sum_j x_{ijk}\beta_j\right)}{1+\exp\left(\beta_{0k} + \sum_j x_{ijk}\beta_j\right)} \qquad \text{(Modelo 3)}$$

p_{ik} es la probabilidad del evento del individuo "i" en la provincia "k", x_{ijk} es el valor de la covariable "j" asociada al individuo "i" en la provincia "k", β_{0k} es el intercepto aleatorio asociado a la provincia "k"

- **Modelos para datos a nivel ecológico**

4) Modelo de Regresión Logística, con agrupación a nivel de Provincias

$$\ln\left(\frac{p_k}{1-p_k}\right)=\beta_0+\sum_j x_{jk}\beta_j+\varepsilon_k \qquad \text{(Modelo 4)}$$

p_k es la proporción del evento observado en la provincia "k", x_{jk} es el valor de la covariable "j" asociada a la provincia "k" (para variables continuas se utiliza el promedio del agrupamiento, para variables categóricas se utiliza el porcentaje), ε_k error aleatorio asociado a la provincia "k"

5) Modelo Poisson, con agrupación a nivel de Provincia, offset asociado al número de eventos esperados para el total del país aca lo que se modela es la tasa del evento (mk/ek)

$$E\left(m_k\right)=\exp\left(\ln e_k+\beta_0+\sum_j x_{jk}\beta_j\right) \qquad \text{(Modelo 5)}$$

m_k es la frecuencia del evento en la provincia "k", e_k frecuencia esperada en la provincia "k" (tomando como población de referencia la de todo el país), x_{jk} es el valor de la covariable "j" asociada a la provincia "k"

6) Modelo de regresión Logística, con agrupación a nivel provincia y estrato (combinación de edad y sexo)

$$\ln\left(\frac{p_{ks}}{1-p_{ks}}\right)=\beta_0+\sum_j x_{jks}\beta_j+\varepsilon_{ks} \qquad \text{(Modelo 6)}$$

ε_{ks} error aleatorio asociado a la provincia "k" en el estrato "s", x_{jks} es el valor de la covariable "j" asociada a la provincia "k" en el estrato "s"

7) Modelo Poisson, con agrupación a nivel provincia y estrato, con offset asociado al número de eventos esperados para el total del país por estrato

$$E\left(m_{ks}\right)=\exp\left(\ln e_{ks}+\beta_0+\sum_j x_{jks}\beta_j\right) \qquad \text{(Modelo 7)}$$

m_{ks} es la frecuencia de desarrollar el evento en la provincia "k" en el estrato "s", x_{jks} es el valor de la covariable "j" asociada a la provincia "k" en el estrato "s", e_{ks} frecuencia esperada en la provincia "k", estrato "s" (tomando como población de referencia la de todo el país en el estrato "s").

Resultados

A fin de comparar el desempeño de los distintos modelos para datos a nivel ecológico, se utilizó una base de datos a nivel individual. En primer lugar se estudió la relación entre obesidad y los factores de riesgo a nivel individual, una vez seleccionado el mejor modelo, se tomaron estos resultados como “gold standard” (o referencia). Luego se forzó el agrupamiento a nivel provincia y a nivel estrato, utilizando porcentajes para variables categóricas y promedios para variables continuas. Finalmente se ajustaron modelos con estos agrupamientos y se compararon los resultados que de ellos proceden con los obtenidos a nivel individual.

Resultados del análisis a nivel individual

Se ajustaron modelos con y sin el efecto fijo asociado a la provincia y luego incluyendo dicho efecto como aleatorio. En la Tabla 1 se muestran los resultados de los tres modelos postulados a nivel individual. Tanto la estimación del efecto aleatorio asociado a la variabilidad entre provincias como la evaluación del efecto de la provincia en el modelo de efectos fijos, indican que la pertenencia a una provincia no aporta significativamente a la explicación de la obesidad. Esto puede deberse a la gran extensión de las provincias y a que la variación es mayor dentro de las provincias que entre ellas.

Tabla 1. Resultados del ajuste de modelos a nivel individual

Variable	Modelo de efectos fijos (sin incluir el efecto provincia)			Modelo de efectos fijos (incluyendo el efecto provincia)			Modelo intercepto aleatorio		
	Estimación	Error Estándar	p-valor	Estimación	Error Estándar	p-valor	Estimación	Error Estándar	p-valor
Intercepto	-1,215	0,091	<,0001	-1,049	0,126	<,0001	-1,218	0,098	<,0001
Edad	-0,008	0,001	<,0001	-0,007	0,001	<,0001	-0,007	0,001	<,0001
Sexo Masculino	0,358	0,036	<,0001	0,360	0,037	<,0001	0,359	0,036	<,0001
Cobertura de salud	-0,062	0,047	0,1886	-0,085	0,047	0,0716	-0,080	0,047	0,092
Consumo de Tabaco(1): Fumador	-0,073	0,046	0,1155	-0,086	0,046	0,0647	-0,082	0,046	0,077
Consumo de Tabaco(1): Ex Fumador	0,198	0,044	<,0001	0,192	0,045	<,0001	0,194	0,044	<,0001
Consumo de frutas y verduras	0,007	0,078	0,9271	0,012	0,078	0,8813	0,010	0,078	0,900
Actividad física(2): Intenso	-0,411	0,060	<,0001	-0,426	0,060	<,0001	-0,421	0,060	<,0001
Actividad física(2): Moderado	-0,194	0,040	<,0001	-0,197	0,040	<,0001	-0,195	0,040	<,0001
Instrucción(3): primario completo	0,100	0,059	0,0905	0,107	0,059	0,0702	0,105	0,059	0,075
Instrucción(3):secundario completo o más	-0,240	0,061	<,0001	-0,230	0,062	0,0002	-0,232	0,061	0,0002
Colesterol	0,278	0,038	<,0001	0,275	0,038	<,0001	0,276	0,038	<,0001
Diabetes	0,548	0,046	<,0001	0,546	0,046	<,0001	0,546	0,046	<,0001
Hipertensión	0,639	0,038	<,0001	0,638	0,039	<,0001	0,637	0,038	<,0001

(1) Ref: No fumador; (2) Ref: Bajo; (3) Ref: primario incompleto

Se toma entonces el modelo sin efecto provincia como el modelo individual contra el cual comparar los modelos agrupados (modelo de referencia). En base al modelo seleccionado los individuos de sexo masculino tienen mayor riesgo de desarrollar obesidad, así como los que son hipertensos, diabéticos y con colesterol. Los que tienen actividad física intensa o moderada tienen menor chance de padecer obesidad que los que tienen actividad física baja.

Resultados del análisis a nivel ecológico

En la Tabla 2 se pueden observar los resultados de los modelos (Modelo 4) y (Modelo 5) ajustados para datos agrupados a nivel de provincias. Es decir, las unidades de observación son las 24 provincias, tanto para la variable respuesta (porcentaje de obesos) como para las explicativas (porcentaje de hipertensos en la provincia, edad promedio, etc).

Tabla 2. Estimación del modelo con datos agrupados a nivel provincia

Variable	Modelo de Regresión Logística		Modelo Poisson	
	Estimación	Significativo	Estimación	Significativo
Intercepto	-0,864		-0,575	
Edad	-0,047		-0,031	SI
Sexo Masculino	-0,023		-0,006	
Cobertura de salud	0,009		0,006	
Consumo de Tabaco(1): Fumador	0,059	SI	0,044	SI
Consumo de Tabaco(1): Ex Fumador	-0,033		-0,021	
Consumo de frutas y verduras	0,032		0,016	
Actividad física(2): Intenso	-0,003		-0,003	
Actividad física(2): Moderado	0,016		0,013	
Instrucción(3): primario completo	-0,007		-0,001	
Instrucción(3):secundario completo o más	-0,007		-0,001	
Colesterol	0,004		-0,001	
Diabetes	0,052		0,029	
Hipertensión	0,014		0,015	

(1) Ref: No fumador; (2) Ref: Bajo; (3) Ref: primario incompleto

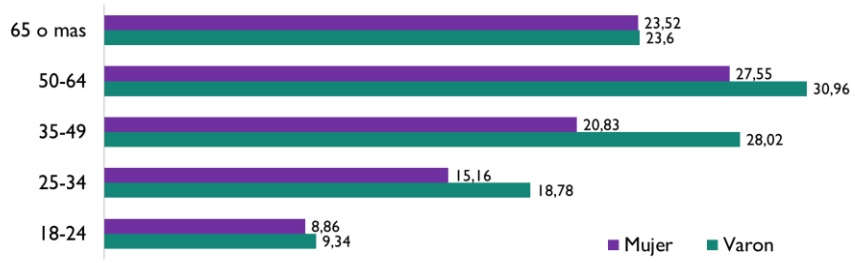
A partir de la Tabla 2 se concluye que en poblaciones con alto porcentaje de hombres se espera un menor porcentaje de obsesos. El porcentaje de individuos con colesterol, diabetes o hipertensión no se ve asociado al porcentaje de obsesos.

Estas conclusiones difieren de las obtenidas en el modelo a nivel individual, lo cual remarca el problema de la falacia ecológica.

A fin de tener más unidades de observación, y que estas sean más heterogéneas, se construyen 10 estratos dentro de cada provincia, definidos por grupos de edad (18-24; 25-34; 35-49; 50-64 y 65 o más) y sexo. Este nuevo agrupamiento permite trabajar con 240 unidades de observación.

En la Figura 1 se muestra el porcentaje de obesos en los distintos estratos a través de todas las provincias.

Figura 1: Distribución de la obesidad por sexo y grupo etario



En la Tabla 3 se presentan los resultados obtenidos para los ajustes de los modelos (Modelo 6 y (Modelo 7) con datos agrupados a nivel de estrato.

Tabla 3. Estimación del modelo con datos agrupados a nivel estrato y provincia

Variable	**Modelo de Regresión Logística**		**Modelo Poisson**	
	Estimación	Significativo	Estimación	Significativo
Intercepto	-4,537	SI	-0,606	
Edad	0,058	SI	-0,016	SI
Sexo Masculino	-0,114		-0,02	
Cobertura de salud	-0,029	SI	0,002	
Consumo de Tabaco(1): Fumador	0,018		0,004	
Consumo de Tabaco(1): Ex Fumador	0,003		0,005	SI
Consumo de frutas y verduras	-0,016		-0,003	
Actividad física(2): Intenso	0,032	SI	0,003	
Actividad física(2): Moderado	-0,02	SI	0,004	
Instrucción(3): primario completo	0,012		0,006	
Instrucción(3):secundario completo o más	0,028	SI	0,004	
Colesterol	-0,001		0,002	
Diabetes	0,02		0,016	SI
Hipertensión	0,009		0,007	SI

La utilización de datos a nivel estratos no minimiza sustancialmente el sesgo atribuido a este agrupamiento.

Conclusiones

Este trabajo pretende, empíricamente, mostrar los problemas ocasionados por la falacia ecológica al momento de interpretar asociaciones.

El modelo individual presentado se considera el “gold estándar”. Ninguno de los modelos propuestos a nivel agrupado logran reproducir las conclusiones obtenidas a nivel individual. La propuesta de utilizar estratos para mejorar las estimaciones tampoco mejora los resultados.

Futuras líneas de investigación deberían dirigirse a: trabajar con dominios más pequeños, combinar diferentes fuentes de datos o trabajar con datos agregados e individuales simultáneamente.

REFERENCIAS BIBLIOGRÁFICAS

Alan Agresti, Foundations of Linear and Generalized Linear Models. Ed. Wiley. 2015

Jackson CH, Best NG, Richardson S. Improving ecological inference using individual-level data. Statistics in Medicine, 2006, 25(12): 2136-2159.

Jackson CH, Best NG, Richardson S. Hierarchical related regression for combining aggregate and survey data in studies of socio-economic disease risk factors, 2008, Journal of the Royal Statistical Society, Series A, 171(1):159-178.

Lancaster G, Green M, Lane S.”Reducing Bias in ecological studies: an evaluation of different methodologies” J.R.Statis.Soc., 2006, 169, Part 4, pp681-700

NIH Public Access. Spatial Aggregation and the Ecological Fallacy. Chapman Hall CRC Handb Mod Stat Methods. ;2010:541-558.

Kutner M, Nachtsheim C, Neter J, Li W, Applied Linear Statistical Models. 5th Edition, Mc Graw-Hill. 2005

FUENTES

Tercera Encuesta Nacional de Factores de Riesgo para Enfermedades No Transmisibles. Argentina 2013. Instituto Nacional de Estadística y Censos.

https://www.indec.gob.ar/nivel4_default.asp?id_tema_1=4&id_tema_2=32&id_tema_3=68 (Consulta: 15/09/2017)