Leticia Hachuel
Gabriela Boggio
Nora Arnesi
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

USO DE MODELOS MARGINALES PARA EL CÁLCULO DE RAZONES DE MORTALIDAD ESTANDARIZADAS*

1. INTRODUCCIÓN

El conocimiento del número de *casos* de algún evento en particular, presente en una población, tiene de por sí poca utilidad para los epidemiólogos o demógrafos si no se lo relaciona con la población de la cual proceden los *casos*. Esta relación se establece generalmente a través de la construcción de tasas.

El término tasa se usa en muchos campos y su significado no es consistente en todos ellos. En demografía, las tasas se definen comúnmente como tasas de *ocurrencia/exposición* (Preston et al., 2001). El numerador de este tipo de tasas contabiliza el número de ocurrencias de un evento de interés, mientras que el denominador combina dos factores: el número de personas en la población y la longitud del tiempo que enmarca el estudio.

En los últimos años se ha intentado mostrar la distribución geográfica de la ocurrencia de enfermedades tales como el cáncer para una región o país particular. Los mapas que muestran esta distribución generalmente se realizan en base a tasas estandarizadas relativas a cada región utilizando las denominadas Razones de Mortalidad Estandarizada (RME). Éstas se construyen comparando el número observado de casos en la población bajo estudio con el valor esperado de casos y generalmente se expresan en porcentajes.

En este trabajo se presenta el cálculo del número esperado de *casos* a partir del ajuste de un modelo de regresión Poisson que considera efectos de los diferentes grupos de edad.

Esta metodología se aplica para el estudio de la mortalidad por cáncer de mama en la provincia de Santa Fe en el período 2000 -2005. El modelo propuesto permite el cálculo de las Razones de Mortalidad Estandarizadas a nivel departamental, controlando el impacto que pueden producir las diferencias en las estructuras por edad de las poblaciones de cada departamento.

En una segunda etapa se realizará el mapeo de estas RME previo control estadístico de la heterogeneidad del tamaño poblacional de los diferentes departamentos de la provincia de Santa Fe.

* Este trabajo se realizó en el marco del proyecto "Modelos para datos multicategóricos y de conteo correlacionados". En el participó, en calidad de auxiliar de investigación, la alumna de la carrera Licenciatura en Estadística Erika Schmidt Strano.

2. LOS DATOS

A partir de la información que contienen los Registros de Estadísticas Vitales, organizada y publicada por el Ministerio de Salud de la Nación sobre la mortalidad en Argentina según causas de muerte (Codificación Diagnóstica Internacional CIE10) se obtiene para cada grupo de edad quinquenal el *número de muertes por tumor maligno de mama en la mujer* en la provincia de Santa Fe desagregado a nivel departamental para los trienios 2000-2002 y 2003-2005. El grupo en estudio se limita a las mujeres de 25 años y más dado que la mortalidad por tumor maligno de mama para las más jóvenes es poco frecuente. De esta forma se dispone de una base que contiene la información correspondiente a los 19 departamentos de la provincia para 11 grupos etáreos, los primeros diez quinquenales y el último reúne los casos acontecidos en las mujeres de 75 años y más.

Las correspondientes personas-año asociadas a cada departamento y grupo etáreo se obtienen a partir de información publicada por el Instituto Nacional de Estadísticas y Censos (INDEC).

3. CÁLCULO DE LAS RAZONES DE MORTALIDAD ESTANDARIZADAS

Las RME se definen como el cociente entre el número de muertes observadas y el número de muertes esperadas en cada área en estudio, en este caso los departamentos de la provincia de Santa Fe.

La obtención de dichos valores esperados se realiza mediante el ajuste de un modelo Poisson marginal que permite controlar la estructura de edad de los diferentes departamentos como así también la posibilidad de respuestas semejantes dentro de cada departamento. El mismo se formaliza:

$$\log \frac{\mu_{ij}}{P_{ij}} = \beta_j , \quad i=1,\ldots,19 \quad \text{y } j=1,\ldots,11$$

donde μ_{ij} representa el valor esperado del número de muertes, Y_{ij}, en el i-ésimo departamento para el j-ésimo grupo de edad , P_{ij}, el número de personas en el i-ésimo departamento para el j-ésimo grupo de edad y los $\{\beta_j\}$ son los parámetros asociados a los diferentes grupos de edad. Para tener en cuenta la posible asociación intra-departamento en relación al número de muertes se considera:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \alpha \qquad j = 1,...t = 11 \quad k = 1,....,t = 11$$

El enfoque de estimación usual para este tipo de modelos es el denominado Ecuaciones de Estimación Generalizadas (GEE), propuesto por Liang y Zeger (1986). Una breve descripción de esta metodología se encuentra en el anexo.

Una vez ajustado el modelo, el número esperado de muertes en el i-ésimo departamento y j-ésimo grupo de edad se calcula resolviendo $\hat{\mu}_{ij} = P_{ij} \exp(\hat{\beta}_j)$.

Finalmente, la suma de estas estimaciones a través de los 11 grupos de edad en cada departamento proporciona el denominador de las RME, es decir el número de muertes esperado para cada de ellos.

4. RESULTADOS

Las RME para los diferentes departamentos de la provincia de Santa Fe para los dos trienios estudiados se presentan en Tabla 1.

Tabla 1: Razones de Mortalidad Estandarizadas para los departamentos de la provincia de Santa Fe en los trienios 00-02 y 03-05

Departamento	Trienio 2000-2002	Trienio 2003-2005
Garay	42.52	115.97
Gral Obligado	56.09	100.90
Iriondo	81.28	110.76
Gral Lopez	89.99	113.06
Constitución	90.41	107.89
Castellanos	75.22	83.25
San Javier	68.49	70.04
La Capital	111.63	109.16
Rosario	110.76	107.05
San Martín	106.81	101.76
Caseros	96.94	91.87
San Cristóbal	76.91	66.85
San Lorenzo	87.49	75.78
Las Colonias	85.67	70.18
9 de Julio	69.55	49.21
Vera	67.73	47.38
Belgrano	121.57	95.65
San Justo	105.45	78.46
San Jerónimo	110.39	77.76

A continuación se presenta un diagrama de caja comparativo entre las RME observadas en el trienio 00-02 y el trienio 03-05, como así también una tabla con las respectivas medidas resumen (Figura 1, Tabla 2).

Figura 1: Razones de Mortalidad Estandarizadas
para los trienios 00-02 y 03-05

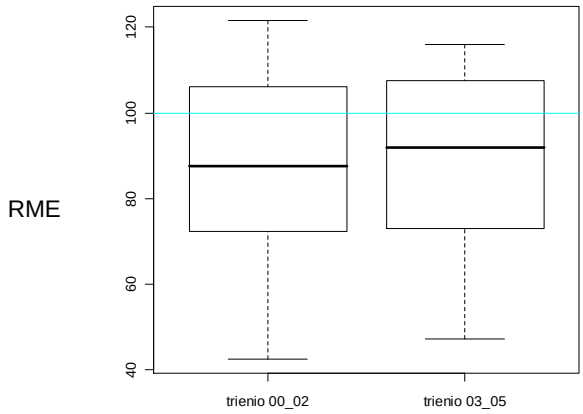


Tabla 2: Medidas resumen de RME para los
trienios 00-02 y 03-05

	Trienio 00-02	**Trienio 03-05**
Mínimo	42.52	47.38
Primer cuartil	72.39	72.98
Mediana	87.49	91.87
Tercer cuartil	106.13	107.47
Máximo	121.57	115.97

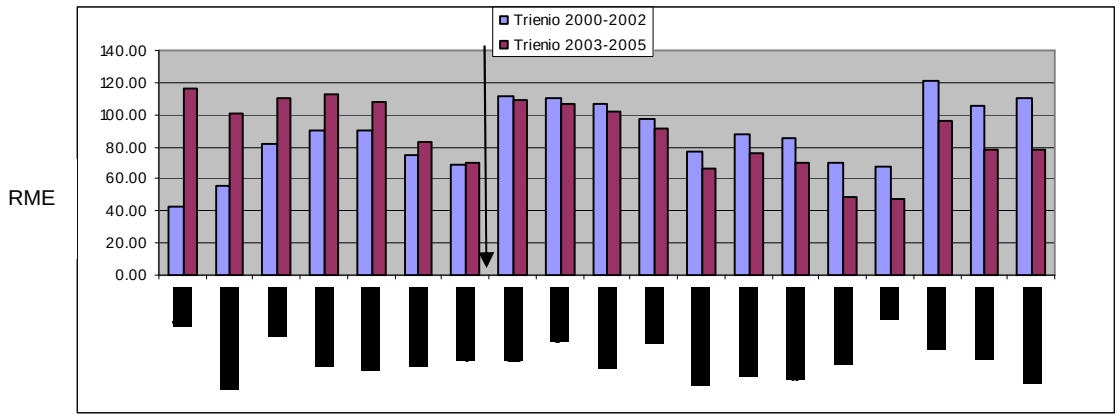
El diagrama de caja precedente muestra que, en ambos trienios, los valores de las RME están más concentrados en el 25% de los valores mayores que en el 25% de los valores más pequeños, los cuales se encuentran muy dispersos.

Si se tiene en cuenta que las RME se construyen como el cociente, expresado en porcentaje, entre el número de muertes observadas en cada departamento y el número de muertes esperadas, un valor igual a 100% corresponde a una situación en la cual las muertes observadas y esperadas coinciden. Por otro lado un valor superior a 100% indicaría una situación en la cual las muertes realmente observadas superan a las esperadas. Por tal motivo en el diagrama se destaca la línea del 100% y se observa que los porcentajes que superan dicha línea son similares en los dos trienios analizados, 32% para el primer trienio y 42% para el segundo. Si bien el porcentaje es mayor para el segundo trienio, hay que destacar que en este período los valores se encuentran más concentrados con valor máximo más bajo que el máximo presentado en el primer trienio (Tabla 2).

Además es posible apreciar que el 50% de los departamentos presenta una RME inferior o igual a 87.49% en el primer trienio, e inferior o igual a 91.87% en el segundo trienio, es decir valores muy similares en ambos períodos.

La Figura Nº 2 presenta los valores de las RME en cada departamento para los dos trienios en estudio.

Gráfico 2: Razón de Mortalidad Estandarizada por departamento para los trienios 00-02 y 03-05



Los departamentos de Garay, Gral. Obligado, Iriondo, Gral. López, Constitución, Castellanos y San Javier presentan una RME inferior en el trienio 00-02 con respecto al trienio 03-05, con la particularidad que en el primer trienio todas las RME están por debajo del 100%. En cambio, en el segundo trienio, las RME de la mayoría de esos departamentos superan dicho valor y sólo Castellanos y San Javier se mantienen por debajo del 100%.

Resulta llamativo, además, el aumento brusco en la RME del departamento de Garay al pasar de un trienio al siguiente. Dicho departamento según información proporcionada por INDEC-IPEC relevada en el Censo de Población y Viviendas 2001, registra el mayor porcentaje de hogares con todos los integrantes sin obra social y/o plan médico o mutual (53.3%) lo cual podría explicar en cierta medida la falta de controles periódicos para detectar el cáncer en estadíos tempranos. Sin embargo, el departamento San Javier que también presenta un porcentaje alto de hogares sin cobertura médica (50,4%), no ha registrado un cambio tan notorio en su RME.

Por otro lado se puede observar que en 12 de los 19 departamentos las RME disminuyen al pasar de un trienio al siguiente. Entre ellos se encuentran algunos con cambios muy pequeños, como es el caso de La Capital, Rosario, San Martín y Caseros, y otros 8 departamentos con cambios más notorios, destacándose San Jerónimo que pasó de una RME de 110.39% a 77.76%.

Posibles explicaciones a estos cambios en magnitud y sentido entre un período y otro necesitan evidentemente tener en cuenta otro tipo de información, estrategia que está siendo llevada a cabo para la posterior construcción del mapeo de las RME.

5. CONSIDERACIONES FINALES

En este trabajo se presenta un procedimiento estadístico para la estimación simultánea de las RME. El cálculo de estas razones en los dos trienios considerados constituye una primera etapa para el correspondiente mapeo. Sin embargo, para la realización de esta representación gráfica conviene tener en cuenta la heterogeneidad del tamaño poblacional de los departamentos, ya que seguramente implica diferentes niveles de precisión en las estimaciones de las RME. Por tal razón se está trabajando en la estimación suavizada de las mismas a partir del ajuste de modelos que incluyan efectos aleatorios y covariables que ayuden a explicar diferencias entre los departamentos. Estos modelos permiten fortalecer

las estimaciones de las RME a nivel departamental haciendo uso de la información conjunta de todos los departamentos de la provincia (Breslow y Clayton, 1993).

6. REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. John Wiley & Sons.

Breslow, N.; Clayton, D. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association,* Vol.88 (421).

Fitzmaurice, G.; Laird, N.; Ware, J. (2004). *Applied longitudinal analysis*. John Wiley & Sons.

INDEC. Censo Nacional de Población, Hogares y Viviendas del año 2001. *www.indec.mecon.gov.ar*, junio 2008.

INDEC. Estimaciones y proyecciones de población total del país. 1950-2015. Serie Análisis Demográfico, 30. *www.indec.mecon.gov.ar*, junio 2008.

Ministerio de Salud. Presidencia de la Nación. *www.msal.gov.ar*, junio 2008.

Preston, S; Heuveline, P; Guillot, M. (2003). *Demography. Measuring and Modeling Population Processes.* Blackwell Publishers.

Song, P. X. K. (2007) *Correlated data analisys: modeling, analytics and applications*. Springer, New York.

ANEXO

Modelo Poisson marginal

Si se dispone de n grupos independientes de t_i observaciones cada uno $(N = \sum_{i=1}^{n} t_i)$ y donde esas t_i observaciones están correlacionadas dentro de cada uno de los grupos, un modelo apropiado es el denominado *modelo marginal*, el cual constituye una generalización de los modelos lineales generalizados para el tratamiento de datos correlacionados (Song, 2007; Fitzmaurice et al., 2004; Agresti, 2002).

Los modelos marginales no requieren supuestos distribucionales para el vector de respuestas $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{it_i})'$ ya que tratan específicamente con los momentos correspondientes a la media y las correlaciones. Ello conduce a un método de estimación conocido como Ecuaciones de Estimación Generalizadas.

Sean $\boldsymbol{\beta} = (\beta_0,\beta_1,..,\beta_p)$ y $\mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2}, ..., x_{ijp})'$. Sea $\boldsymbol{\beta}'\mathbf{x}_{ij}$ el predictor lineal que se relaciona con las esperanzas marginales a través de la expresión $g(\mu_{ij}) = \boldsymbol{\beta}'\mathbf{x}_{ij}$, siendo $\mu_i = (\mu_{i1},....,\mu_{it_i})'$ la esperanza marginal de $\mathbf{Y}_i$. La consideración de $g(\mu_{ij}) = \log\mu_{ij}$ conduce al modelo Poisson marginal que puede especificarse de la siguiente forma:

1- La media de Y_{ij} se relaciona con las covariables mediante el enlace logaritmo:

$$\log\mu_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij}.$$

2- La variancia de cada Y_{ij}, dados los valores de las covariables, depende de la respuesta media:

$$\text{Var}(Y_{ij}) = \mu_{ij}\,.$$

3- La asociación condicional intra-grupo del vector de respuestas repetidas dadas las covariables se considera función de un conjunto adicional de parámetros de asociación, $\boldsymbol{\alpha}$:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \alpha_{jk} \quad j = 1,...t_i \quad k = 1,....,t_i\,.$$

Liang y Zeger (1986) introdujeron un método para incorporar la correlación intra-grupo, generan las denominadas Ecuaciones de Estimación Generalizadas (GEE):

$$U(\boldsymbol{\beta};\mathbf{R}) = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}\,, \quad \text{donde:}$$

$\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}/\phi$, con $\mathbf{R}_i(\boldsymbol{\alpha})$ matriz $t_i x t_i$ de correlación de trabajo que modela la correlación entre las t_i observaciones como una función de los parámetros del vector $\boldsymbol{\alpha}$, y ϕ es un parámetro de escala,

$\mathbf{D}_i$ es una matriz de dimensión t_i x (p+1) que tiene como elemento (j,k) a $\partial\mu_{ij}/\partial\beta_k$ y

$\mathbf{A}_i$ es una matriz diaglonal de dimensión $t_i x t_i$ que tiene como elemento (j,j) la variancia de Y_{ij}, μ_{ij}.

Estas ecuaciones constituyen el pivote para la estimación de los parámetros del modelo marginal.

El método GEE produce estimadores de los parámetros $\boldsymbol{\beta}$ asintóticamente normales y consistentes (suponiendo condiciones débiles de regularidad y la correcta especificación de la función de la media) cuando $\boldsymbol{\alpha}$ se reemplaza por una estimación $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}_{GEE})$ que converge a $\boldsymbol{\alpha}$ suficientemente rápido, aún si el modelo para la correlación de trabajo y/o la matriz $\mathbf{A}_i$ son incorrectas (Liang y Zeger, 1986). El propósito de la estimación original fue iterar entre la estimación por momentos de ($\boldsymbol{\alpha}$, ϕ) para $\boldsymbol{\beta}_{GEE}$ fijo, y, mediante un método modificado de Newton-Raphson, resolver las GEE para los $\hat{\boldsymbol{\beta}}_{GEE}$, considerando $\boldsymbol{\alpha}$ y ϕ fijos. Liang y Zeger (1986) usaron funciones simples de los residuos Pearson para estimar los parámetros de correlación en $\boldsymbol{\alpha}$**:**

$$\hat{r}_{ij} = \{y_{ij} - \hat{\mu}_{ij}\}/\sqrt{\hat{\mu}_{ij}}$$

y sugirieron como estimación de ϕ a:

$$\hat{\phi}^{-1} = \sum_{i=1}^{n}\sum_{j=1}^{t_i} \hat{r}_{ij}^2 \Big/ \left(\sum_{i=1}^{n} t_i - p\right).$$

La forma particular del estimador $\hat{\boldsymbol{\alpha}}$ depende de la estructura de la matriz de correlación de trabajo. Para su presentación, a los efectos de simplificar, se supone que todas las unidades presentan t_i=t observaciones, i=1,.. ,n.

Se han sugerido muchas opciones para la estructura de correlación de trabajo. El modelo de trabajo de *independencia*, con **R** igual a la matriz identidad **R**=**I,** adopta la suposición que las observaciones dentro de cada grupo son independientes:

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

Un modelo de correlación de trabajo de uso frecuente es el denominado *constante* ("exchangeable"), el cual asume que la correlación entre dos observaciones cualesquiera de la misma unidad es fija, es decir, $R_{jj'}$=α, para j≠j'. Esta estructura de correlación podría ser escrita en términos de un sólo parámetro de correlación simple, $0 < \alpha < 1$:

$$\mathbf{R} = \begin{pmatrix} 1 & \alpha & \ldots & \alpha \\ \alpha & 1 & \ldots & \alpha \\ \ldots & \ldots & \ldots & \ldots \\ \alpha & \ldots & \alpha & 1 \end{pmatrix}.$$