Creación de indicadores alternativos para la vigilancia en salud oral, mediante Regresión Beta

Ramón Alvarez Vaz-Elena Vernazza[1]

[1]Instituto de Estadística. Facultad de Ciencias Económicas y Administración

Resumen

En el ámbito de la Salud Pública, existe la necesidad de conocer en profundidad las características de las poblaciones y los problemas de salud. De esa manera se puede intervenir para mejorarlos. Es fundamental tener una idea de la situación de partida y para eso se recurre a las fuentes de datos existentes, fundamentales en los sistemas de vigilancia epidemiológica.
Sin embargo pueden existir limitaciones en los indicadores generalmente utilizados en la epidemiología y salud pública, ya que muchas veces no toman en cuenta la estructurada multivariada de la información o si la toman, lo hacen a través de algoritmos de cálculo que generan indicadores univariados para ganar en simplicidad, y no miden por lo tanto, correctamente los fenómenos bajo estudio.
Se propone construir un conjunto de indicadores alternativos y complementarios a los que ya existen en salud oral. Se reformularán los índices recomendados de la Organización Mundial de la Salud (CPO,SIC); en el estudio de la salud oral de la población habitualmente se usa el CPO (número total de piezas cariadas, perdidas y obturadas).
La aplicación se hace con información proveniente del relevamiento en necesidades de tratamiento y demanda de servicios de salud bucal, de la población de Trabajo por Uruguay (TPU) a agosto de 2007.
Se presentan las limitaciones que tiene trabajar con un *índice agregado univariado*. Como alternativa se propone modelar proporciones construídas a partir del CPO, para lo que se utilizan distintos tipos de *modelos lineales generalizados (GLM)* adaptados para modelar proporciones. Se utilizan modelos de *regresión beta* (mediante una reparametrización adecuada), propuesta por Ferrari y Cribari-Neto (2004) para modelar variables de respuesta continua a valores en el intervalo $(0,1)$. En este caso las variables explicativas tienen que ver con aspectos sociodemográficos individuales (edad, sexo, educación), de contexto (región, barrio) y con la historia de salud bucal de cada individuo (el motivo de su consulta, la cantidad de prótesis, el tiempo sin concurrir al dentista, etc)

Presentación del problema

Indice CPO:

- índice *unidimensional* que cuenta el número de dientes cariados (C), perdidos (P) y obturados (O)
- determina la historia de salud, medido a través de la *caries* de un conjunto de individuos
- valores bajos indican un buen 'status' de salud oral (poca hstoria de enfermedad)
- para una persona en particular se puede evaluar el estado de las piezas a través del índice que se detalla en la siguiente ecuación:

$$CPO^g_{i,j,k} = \sum_j^n C^g_{i,j,k} + \sum_j^n P^g_{i,j,k} + \sum_j^n O^g_{i,j,k} \qquad (1)$$

donde $i = individuo$; $j = diente$; $k = cuadrante$; $g = grupo$ (por ej,sexo)

- Problema de este indicador: Enmascara variablidad de los componentes. Un mismo valor de *CPO* de 12 puede estar indicando situaciones muy diversas, como de una persona con 8 piezas obturadas y 4 con caries, y de otra con 5 cariadas y 7 perdidas.
- Posible solución: utilizar los 3 componentes del CPO por separado, transformado en tasas, o proporciones.

La encuesta

Un equipo de Investigadores de Facultad de Odontología realizó un relevamiento epidemiológico, de necesidades de tratamiento y demanda de servicios de salud bucal para una muestra de la población de TPU. La población de estudio se compone de 1185 personas mayores de 18 años correspondiente al tercer llamado de TPU en Montevideo. Se obtuvo una muestra aleatoria mediante un muestreo multietápico (n=308) de individuos provenientes de tres regiones de Montevideo.
Algunas de las variables estudiadas fueron:

- **subregion** de la ciudad de Montevideo
- **edad** (en tramos)
- **sexo**
- **nivel educativo** (primaria, ciclo básico y bachillerato)
- fecha de la **ultima consulta** y **motivo** de consulta.

Modelos probabilísticos para ajustar tasas

Se reformulan las variables originales (variables de conteo) como tasas o proporciones, relativizando contra diferentes totales: máximo número de piezas(32), número de piezas presente,etc. La ventaja de estas transformaciones es que existen modelos probabilísticos conocidos para trabajar con tasas, proporciones o índices de concentración, de los cuales se conocen muchas características necesarias a la hora de usarlos para hacer *inferencias*, considerando que:

- Las proporciones a estimar están en el rango $(0,1)$.
- No cumplen el supuesto de **Normalidad**.
- Pueden existir asimetrías muy importantes.
- La varianza puede cambiar, lo que obliga a manejar otros modelos

Proporciones a estimar

En este trabajo se relativizan los componentes del CPO convirtiéndolos en tasas o proporciones de la siguiente manera

1. $\frac{\sum O_i}{\sum O_i + \sum C_i}$ nivel de cobertura de la enfermedad previa a la entrada al programa (*prop1*)
2. $\frac{\sum C_i}{\sum O_i + \sum C_i}$ indicador de estadio de la enfermedad en el momento actual (*prop2*)
3. $\frac{\sum P_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$ indicador de necesidad de prótesis en el momento actual (*prop4*)

Formulación del modelo de probabilidad BETA

El modelo de probabilidad Beta se presenta de la siguiente manera:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1, \qquad (2)$$

- con $0 < \mu < 1$, $\phi > 0$. Cribari y Neto [Cribari-Neto and Zeileis, 2010] hacen una reparametrización $\mu = \frac{p}{p+q}$ y $\phi = p + q$.
  Se puede escribir $y \sim \mathcal{B}(\mu, \phi)$. Por lo tanto, $E(y) = \mu$, $VAR(y) = \mu(1-\mu)/(1+\phi)$.
- El parámetro ϕ se conoce como parámetro de precisión, ya que para μ fijo, cuanto más grande es ϕ más pequeña es la varianza de y; ϕ^{-1} es un parámetro de dispersión.

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \qquad (3)$$

Diferentes modelos de distribución BETA

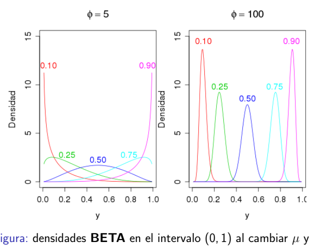

Figura: densidades **BETA** en el intervalo $(0,1)$ al cambiar μ y ϕ

Formulación del modelo predictivo

A partir de los trabajos de Salinas et al [Salinas-Rodríguez et al., 2009] podemos decir que

- Si se tiene $y_1, \dots, y_n$ una muestra aleatoria tal que $y_i \sim \mathcal{B}(\mu_i, \phi)$, $i = 1, \dots, n$.
- El **modelo de regresión BETA (MRB)** es $\mu_i = g^{-1}(x_i^\top \beta)$ donde β, el vector de parámetros de regresión, se estima por máxima verosimilitud (ML).

En varias ocasiones el parámetro de precisión no es constante a través de todas las observaciones, con lo cual es necesario modelarlo, tal cual se hizo con la **media**. En particular $y_i \sim \mathcal{B}(\mu_i, \phi_i)$, $i = 1, \dots, n$, y

$$g_1(\mu_i) = \eta_{1i} = x_i^\top \beta, \; g_2(\phi_i) = \eta_{2i} = z_i^\top \gamma \qquad (4)$$

donde $\beta = (\beta_1, \dots, \beta_k)^\top$, $\gamma = (\gamma_1, \dots, \gamma_h)^\top$, $k + h < n$, son los **coeficientes de regresión** de ambas ecuaciones, η_{1i} and η_{2i} son predictores lineales, x_i y z_i son lo vectores de regresión, los que se estiman por (ML).La función de **link** puede ser **logit, probit,log(log)**

Distribución de los componentes del CPO

La distribución de los componentes del CPO es la que se muestra en la siguiente figura, donde se ve una gran asimetría para el componente de caries y para el componente de obturación, lo que habla de una población muy enferma (enfermedad actual y/o pasada).

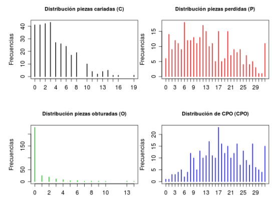

Figura: Densidades empíricas de los **componentes del CPO**

Distribución de los componentes del CPO - Proporciones

Si trabajamos con proporciones, la distribución para cada uno de los componentes, es la siguiente.

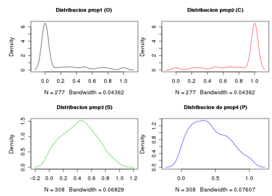

Figura: densidades de los **componentes del CPO** en proporciones

Modelo Regresión Beta (MRB) para C

MRB para la proporción de caries definida como: $prop6 = \frac{\sum C_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$

	Coeficiente	Error Std.	valor t	Pr(>|t|)
(Intercept)	1.1106	0.1213	9.16	0.0000
edad.recDe 35 a 44	-0.4342	0.1113	-3.90	0.0001
edad.recMayor a 45	-0.9559	0.1348	-7.09	0.0000
CPO	0.0303	0.0066	4.59	0.0000
	Coeficiente	Error Std.	valor z	Pr(>|z|)
De Dispersión(ϕ)	7.2103	0.6104	11.81	$< 2e-16$

Cuadro: Modelo De Regresión Beta (MRB) para C vs edad y CPO

¿Cómo seguir?

- Existe el problema que al expresar las tasas (para los componentes del CPO) aparecen 0 y 1, donde la variable $\mathcal{B}$ no está definida aspecto que se resuelve mediante la transformación de *Smithson-Verkuilen* [Verkuilen and Smithson, 2012] $y_i^* = \frac{y_i(n-1)+0{,}5}{n}$.
- Por otra parte hay que trabajar con los modelos que consideran los excesos de 0, que se denominan *zero inflated models* [Ospina and Ferrari, 2012]
- Otro camino es seguir con la extensión de esta clase de modelos **BETA** [Grün et al., 2011], donde se consideran
  - Estimación de mejores modelos mediante árboles de regresión (Combinación de los métodos CART y los (MRB))
  - (MRB) con variable de clase latente que explican mejor la *heterogeneidad*

Referencias Bibliográficas

Cribari-Neto, F. y Zeileis, A. (2010).
Beta regression in r.
Journal of Statistical Software, 34(2):1–24.

Grün, B., Kosmidis, I., y Zeileis, A. (2011).
Extended beta regression in R: Shaken, stirred, mixed, and partitioned.
Working Paper 2011-22, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck.

Ospina, R. y Ferrari, S. L. (2012).
A general class of zero-or-one inflated beta regression models.
Computational Statistics & Data Analysis, Volume 56(Issue 6):1609â1623.

Salinas-Rodríguez, A., Manrique-Espinoza, B., y Sosa-Rubí, S. G. (2009).
Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud.
Salud Pública de México, 51:397 – 406.

Verkuilen, J. y Smithson, M. (2012).
Mixed and mixture regression models for continuous bounded responses using the beta distribution.
Journal of Educational and Behavioral Statistics, 37(1):82–113.