Revista de la Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 10 N°2 julio-diciembre de 2021 • ISSN-e 2357-5749 • Artículo Investigación • Páginas 105 a 125
DOI: https://doi.org/10.15446/rev.fac.cienc.v10n2.80743

MODELOS DE CONTEO ALTERNATIVOS PARA LOS COMPONENTES DEL CPO EN UNA ENCUESTA DE SALUD BUCAL EN MONTEVIDEO, URUGUAY[a]

ALTERNATIVE COUNTING MODELS FOR DFM COMPONENTS IN AN ORAL HEALTH SURVEY IN MONTEVIDEO, URUGUAY

RAMÓN ÁLVAREZ-VAZ [b*], FERNANDO MASSA[c]

Recibido 28-06-2019, aceptado 02-06-2021, versión final 30-06-2021.
Artículo Investigación

RESUMEN: Actualmente existen varios indicadores de las diferentes dimensiones que se determinan a nivel individual en salud oral desde una perspectiva epidemiológica. Dentro de los que corresponden a la patología Caries se consideran entonces los indicadores CPO, ceo, ICDAS entre otros. El CPO es un índice unidimensional que cuenta el número de dientes cariados C, perdidos P y obturados O y cuando debe ser evaluado en un contexto de regresión, es un caso particular de Modelo de Conteo, donde la variable de respuesta refiere al número de veces que un evento ocurre, siendo el evento de conteo la realización de una variable aleatoria no negativa, pudiéndose trabajar con el marco conceptual de la teoría de los Modelos Lineales Generalizados (MLG). En la revisión de la literatura en varios trabajos publicados en revista especializadas de Biomedicina y Epidemiología bien rankeadas no se le presta mucha atención a éstos aspectos, donde no queda muchas veces claro porqué se opta por alternativas al modelo de Poisson, sino que tampoco se trabaja la capacidad de ajuste (ver capacidad predictiva). Los autores muchas veces solamente se dedican a ver las variables y ajustar modelos, resolviéndose por aquellos donde aparecen variables significativas pero que podrían ser muy pobres prediciendo. Este último aspecto es relevante ya que en base a esos modelos los investigadores terminan elaborando teoría para explicar patologías en función de variables que no son buenas predictoras. Por estos motivos en este trabajo se presentan alternativas a los modelos de conteo básicos y se pone un especial énfasis en la capacidad predictiva de los mismos.

PALABRAS CLAVE: CPO; modelos de conteo; modelos hurdle; modelos de Poisson; sobredispersión.

ABSTRACT: Currently, there are several indicators of the different dimensions that are determined at the individual level in oral health from an epidemiological perspective. Among those that correspond to the Caries pathology, the DFM, ceo, ICDAS indicators are considered. DFM is a one-dimensional index that counts the number of teeth decayed D, filled F, missing M and when it should be evaluated in a regression context, is a particular case of a counting model, where the response variable refers to the number of times an event occurs; this counting event is the realization of

[a] Álvarez-Vaz, R. & Massa, F. (2021). Modelos de conteo alternativos para los componentes del CPO en una encuesta de salud bucal en Montevideo, Uruguay. *Rev. Fac. Cienc.*, 10(**2**), 105–125. DOI: https://doi.org/10.15446/rev.fac.cienc.v10n2.80743

[b] Dr. en Ciencias Médicas. Prof. Agregado, Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República,Uruguay

[*] Autor para la correspondencia: ramon.alvarez@fcea.edu.uy.

[c] M. Sc. en Ingeniería Matemática. Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República,Uruguay

a non-negative random variable, being able to work with the conceptual framework of the theory of the Generalized Linear Models (GLM). In the review of the literature in several papers published in well-ranked specialized journals of Biomedicine and Epidemiology, little attention is paid to these aspects,where it is not often clear why the authors opt for alternatives to the Poisson model, and they do not work on the adjustment capacity either. Authors often only look at the variables and adjust models, resolving themselves by those where significant variables appear but which could be very poor predicting. This last aspect is relevant since based on these models the researchers then in the discussion end up elaborating theory to explain pathologies based on variables that are not good predictors.

KEYWORDS: Count models; DMF; hurdle models; overdispersion; Poisson models.

1. INTRODUCCIÓN

Los problemas de salud en la mayor parte del mundo han cambiado, siendo las enfermedades no transmisibles (ENT) las de mayor prevalencia, donde este nuevo patrón global está estrechamente relacionado a los estilos de vida de las sociedades modernas, (Breilh, 2010). Este cambio también impactó en la salud bucal, siendo las enfermedades bucodentales unos de principales problemas en la salud pública debido a su alta prevalencia e incidencia en todas regiones del mundo, (Petersen, 2004).

Antes de pasar a presentar las patologías bucales mas importantes y los índices para evaluarlas es necesario consignar previamente que en Odontología existe una forma de referirse a las piezas dentales, para las que existe cierta numeración coherente también con la medición de patologías de la mucosa. Las piezas se suelen también agrupar en sextantes (inferiores y superiores) o cuadrantes (inferiores y superiores y a su vez izquierdos y derechos). En la Figura 1 denominada Odontograma puede verse la disposición de las piezas en la boca, donde las piezas numeradas del 55 al 75 corresponden a la temporarias.

El concepto actual de Caries dental define a la enfermedad como un proceso dinámico localizado en la superficie dentaria cubierta por bio-película, caracterizado por desequilibrios en los procesos desmineralización-remineralización que ocurren constantemente en la cavidad bucal. A lo largo de un determinado período de tiempo, la predominancia de momentos de pérdida mineral de los tejidos duros del diente (principalmente iones calcio y fosfato) para la bio-película y saliva resulta en el establecimiento de la lesión de Caries (Holst *et al.*, 2001). Algunos componentes del proceso de Caries actúan en la superficie del diente (saliva, bio-película, dieta, acceso al flúor), mientras que existen otro conjunto de factores que determinan el comportamiento de la persona (conocimiento, actitud, ingresos, nivel educativo y socioeconómico), (Fejerscov, 2004). El proceso de la enfermedad debe ser sujeto a un control permanente a lo largo de la vida con el fin de evitar consecuencias irreversibles en etapas posteriores (Maltz & Jardim, 2010).

Existen varios indicadores de las diferentes dimensiones que se determinan a nivel individual en salud oral que pueden ser considerados a nivel colectivo desde una perspectiva epidemiológica.

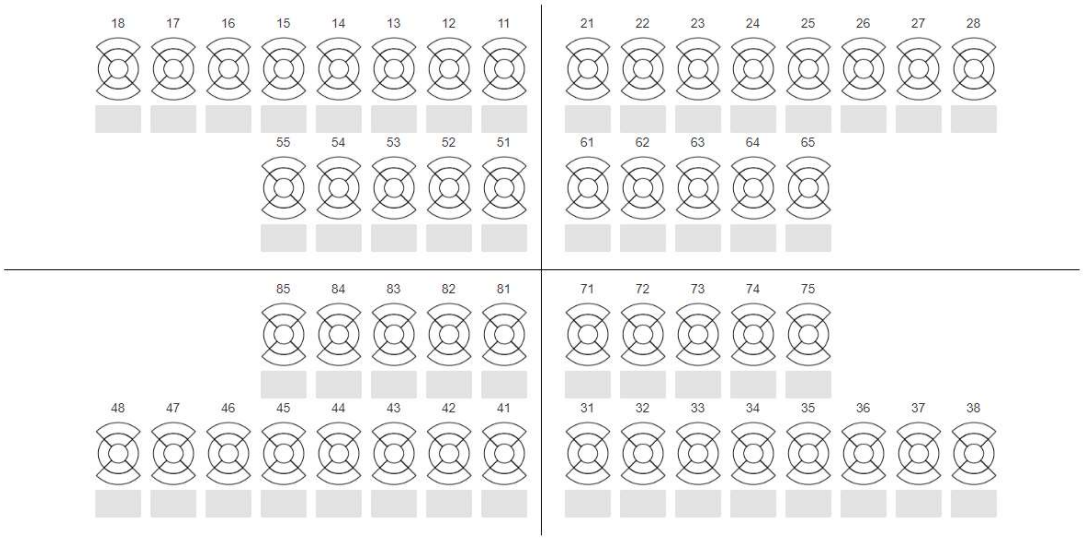

Figura 1: Odontograma, Fuente: Salud dental para todos (https://dtdental.co/que-es-un-odontograma-dental/).

Dentro de los que corresponden a la patología Caries se consideran entonces los indicadores ceo, CPO, ICDAS. De los índices que dan cuenta del estado de las piezas dentales, es necesario hacer definiciones y establecer una nomenclatura de los diferentes unidades de observación que se tomarán de ahora en adelante con la siguiente simbología:

- i individuo que puede estar en $1,\ldots,n$, j que toma valores en $1,\ldots,8$ para el diente, k de $1,\ldots,4$ para el cuadrante, g grupo o subpoblación (se podría tener, por ejemplo subopoblaciones: hombres y mujeres) con valores en $1,\ldots,G$;
- Las piezas dentales $d^g_{i,j,k}$;
- Los cuadrantes $q^g_{i,.,k}$ (formados por piezas);
- Las superficies de cada pieza $s^g_{i,j,l}$

El CPO es un índice *unidimensional* que cuenta el número de dientes cariados C, perdidos P y obturados O. Ha sido utilizado durante mucho tiempo como una forma de determinar la historia de salud, medido a través de la *Caries* de un conjunto de individuos. Los valores bajos de CPO indican un buen ’status’ de salud oral, mostrando que las piezas dentales tienen poca hstoria de enfermedad. Generalmente las personas tienen, salvo excepciones, un total de $28-32$ piezas, repartidas en 4 cuadrantes, 2 inferiores y a su vez izquierdos y derechos, con un total de 7 piezas por cuadrante. Cuando las personas tienen incluso los terceros molares (lo que se habitualmente se llaman ’muelas del juicio’) se puede tener hasta 32 piezas, con un total de 8 por cuadrante. De esta manera, para una persona en particular se puede evaluar el estado de las piezas a través

Tabla 1: Relación entre Media y Varianza para diferentes modelos de Conteo

Modelo	Media	Varianza
Poisson	μ	μ
Binomial Negativa (BN - tipo I)	μ	$\mu(1+\gamma) = \mu+\gamma\mu$
Binomial Negativa (BN - tipo II)	μ	$\mu(1+\gamma\mu) = \mu+\gamma\mu^2$
Binomial Negativa ρ	μ	$\mu(1+\gamma\mu^{\rho}) = \mu+\gamma\mu^{\rho}$
PIG	μ	$\mu(1+\gamma\mu^2) = \mu+\gamma\mu^3$
PG	μ	$\mu(1+\gamma\mu)^2 = \mu+2\gamma\mu^3+\gamma^2\mu^3$

del índice que se detalla en la siguiente ecuación:

$$CPO_i^g = \sum_j^n C_{i,j,k}^g + \sum_j^n P_{i,j,k}^g + \sum_j^n O_{i,j,k}^g. \tag{1}$$

Sin embargo, el primer problema que presenta este indicador es que enmascara toda la variabilidad de las diferentes dimensiones que mide (2 de enfermedad presente (C,P) y 1 de enfermedad pasada pero curada O. Por ejemplo, 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. En ambos casos, los niveles de enfermedad son importantes (tienen 12/28 % de su piezas afectadas, es decir 'no sanas') pero no se sabe si la carga de enfermedad es la misma, ya que las piezas obturadas ponen de manifiesto la enfermedad Caries en el pasado.

2. DIFERENTES DISTRIBUCIONES DE PROBABILIDAD PARA MODELOS DE CONTEO

Casi tan importante como poder modelar adecuadamente los modelos de conteo es fundamental en primer lugar identificar las posibles distribuciones de probabilidad. Para eso, en la Tabla 1 se presentan diferentes alternativas de modelos de probabilidad, en las que se modula la varianza en función de un factor de inflación γ y donde queda por lo tanto determinada una forma de variar la varianza que puede ser lineal como en el caso de la (*BN*- tipo II) y en forma cuadrática para la (*BN*-tipo I).

Para los casos de las *PIG* y *PG* las funciones de varianza son polinomios de grado 3 en μ, (Hilbe, 2011). Se puede comenzar por el caso más sencillo, con el Modelo de Poisson (MPoi), que tiene la siguiente función de cuantía, donde y expresa la variable aleatoria de conteo,

$$f(y;\mu) \quad = \quad \frac{\exp(-\mu)\cdot\mu^y}{y!} \tag{2}$$

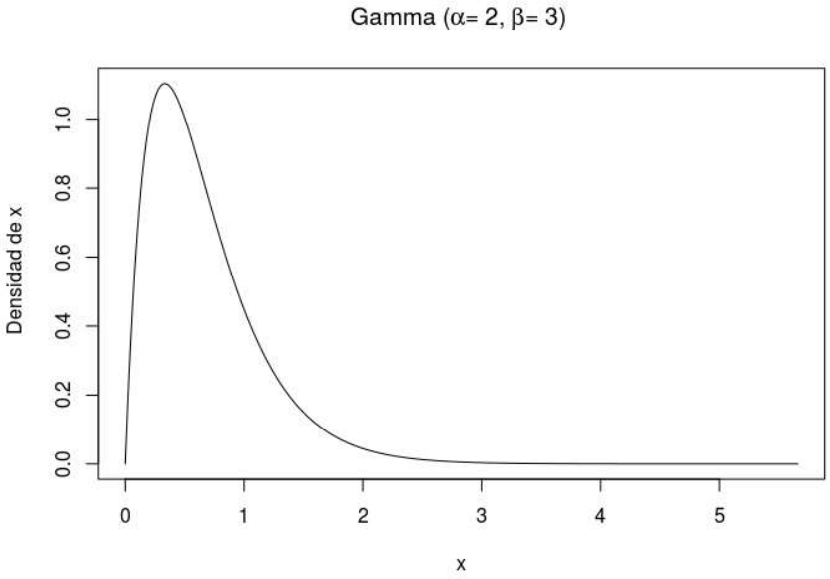

Figura 2: Densidad **Gamma**(α,β) para parámetro μ en distribución *BN*, Fuente: Elaboración propia.

Cuando existe sobredispersión se puede trabajar con un modelo (MBN), que es una mezcla de distribuciones de Poisson, con diferentes μ_i y el proceso de mezcla está dado por una distribución Γ para μ, donde $\mu \sim \text{Gamma}(\alpha,\beta)$, (en realidad es lo que se llama mezcla Poisson-Gamma), es decir:

$$f(y|\mu) = \int_0^\infty \frac{\exp(-\mu)\cdot\mu^y}{y!} f_\Gamma(\mu)d\mu \tag{3}$$

$$\mu \sim \text{Gamma}(\alpha,\beta) \tag{4}$$

en donde el parámetro α controla la forma de la distribución y β su escala. Además su función de densidad está dada por:

$$f(\mu) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu} & \text{si } \mu \geq 0 \quad (\alpha>0,\ \beta>0) \\ 0 & \text{en otro caso} \end{cases} \tag{5}$$

En la Figura 2 puede verse cual una posible forma de variar el parámetro μ, de acuerdo a una distribución $Gamma(\alpha,\beta)$.

Manejando la notación de Hilbe (2011), finalmente la Binomial Negativa (BN) de tipo II se puede expresar como:

$$f(y;\mu,\gamma) = \binom{y+\frac{1}{\gamma}-1}{\frac{1}{\gamma}-1}\left(\frac{1}{1+\gamma\mu}\right)^{\frac{1}{\gamma}}\left(\frac{\gamma\mu}{1+\gamma\mu}\right)^y. \tag{6}$$

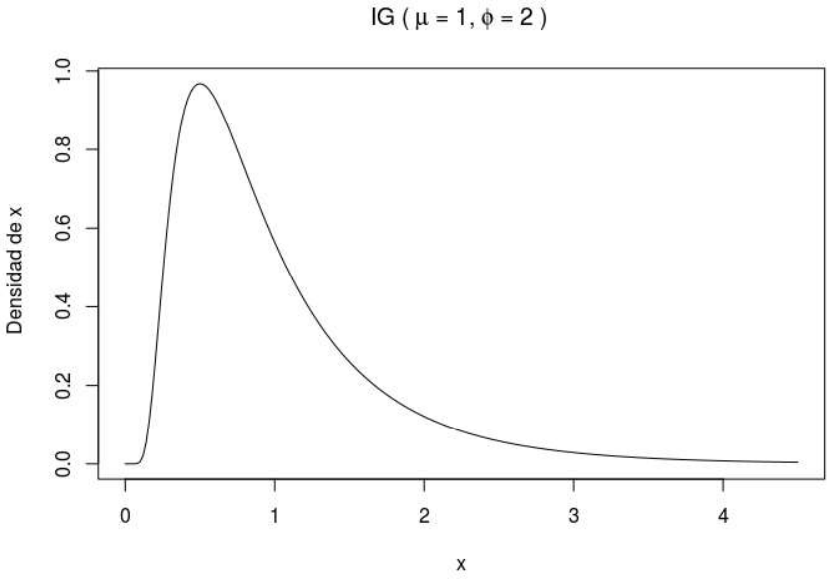

Figura 3: Densidad **IG**(μ,ϕ), para parámetro μ en distribución *PIG*, Fuente: Elaboración propia.

Otra distribución de probabilidad para la forma de variar μ es del tipo Inversa Gaussiana(IGau). Es una distribución muy usada en Análisis de Sobrevida en el campo Médico y Actuarial, (Whitmore, 1975; Whitmore y Yalovsky, 1978), caracterizada por ser asimétrica con parámetros de media μ y precisión ϕ. Esto da origen a la Poisson Inversa Gaussiana (PIG), que se debe interpretar en forma similar al proceso de mezcla de Poisson-Gamma para la BN, con la diferencia que en este caso es una mezcla de Poisson y de Inversa Gaussiana, con $\alpha = \frac{1}{\phi}$, (Giner & Smyth, 2016; Wheeler, 2016).

$$f(y;\mu,\alpha) = \begin{cases} \sqrt{\dfrac{\phi}{2\pi y^3}}\, exp\left\{\dfrac{-\phi(y-\mu)^2}{2\mu^2 y}\right\}, & \text{si } 0 < y < \infty \\ 0 & \text{en otro caso} \end{cases} \tag{7}$$

Cuando se presentan datos que muestran que el parámetro de dispersión γ puede no ser fijo a lo largo de todas las observaciones, como es el caso de la BN y la PIG, es necesario incorporar un parámetro extra ρ que interviene en lo que se conoce como (MBNp).

Si se tiene en cuenta la función de varianza que surge de la Tabla 1, para la BN tipo I y la BN tipo II la diferencia es la siguiente:

$$\begin{cases} \text{Binomial Negativa (BN - tipo I)} & \mu & \mu(1+\gamma\mu) = \mu+\gamma\mu \\ \text{Binomial Negativa (BN - tipo II)} & \mu & \mu(1+\gamma) = \mu+\gamma\mu^2 \\ \text{Binomial Negativa} - \rho & \mu & \mu(1+\gamma\,\mu^{\rho}) = \mu+\gamma\mu^{\rho} \end{cases}$$

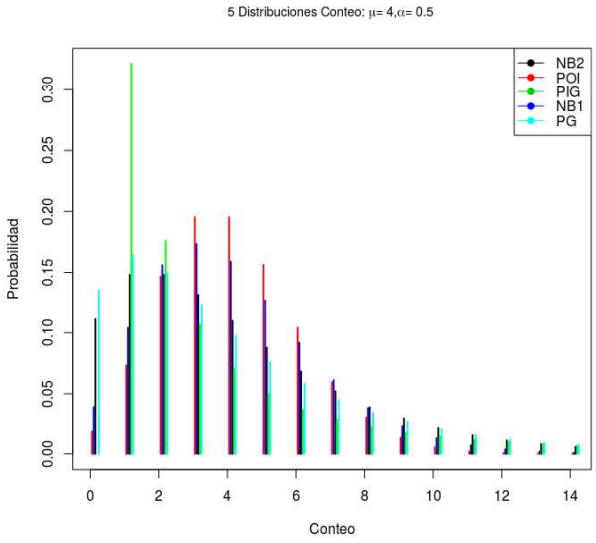

Figura 4: Comparación de diferentes distribuciones de Probabilidad para Modelos de Conteo, Fuente: Elaboración propia.

es decir, una potencia del término en $\gamma\mu$, donde ρ debe estimarse junto con el resto de los parámetros (Hilbe, 2011).

Por último, si bien es poco frecuente encontrar datos con *subdispersión*, existe otra alternativa de distribución de Probabilidad que es la *PG*, que tiene la siguiente expresión, con $\theta > 0$ y $\|\delta| < 1$:

$$f(y;\theta,\delta) = \frac{\theta_1(\theta_1+\delta y)^{y-1}e^{-\theta_1-\delta y}}{y!}, \quad y = 0,1,2,\ldots \tag{8}$$

En Hilbe (2011), el autor presenta un estudio sobre tiempo en días de internación (TDI) de pacientes con patología coronaria que reciben 2 tipos de procedimientos quirúgicos, los que muestran tener una media para TDI de 8.8 y un desvío éstandar de 6.9, que muestra una sobre dispersión de 3.2. Sin embargo de los 3589 observaciones originales se intenta modelar el TDI de los que tienen menos de 8 días, siendo 1982 pacientes que verifican esa restricción mostrando una media de TDI de 4.4, con un desvío estándar de 2.30, con una dispersión de 0.79, lo que indica que no es conveniente en el contexto de regresión usar un MPoi o un MBN, con lo cual una alternativa es precisamente usar la distribución PG.

Es importante entonces, más allá de ver concretamente cada modelo antes planteado, compararlos visualmente, por lo cual para un valor dado de $\mu = 4$ y $\gamma = 0.5$, se presenta como se diferencian entre éstos.
Sin embargo, en la práctica se encuentran datos que presentan otras patologías como son el exceso de ceros o dado el problema, la no presencia de ceros como puede ser por ejemplo: los clásicos estudios de días de internación (donde se puede definir que el mínimo es 1), donde por definición el recorrido está truncado. Para eso se puede recurrir a los modelos que siguen.

2.1. Modelos Hurdle (MH)

Los *MH* o *Hurdle Models*, que podrían considerarse como modelos con *obstáculos*, son aquellos que combinan dos procesos de conteo, uno para los ceros , con $f_{\text{cero}}(y;z,\gamma)$ (censurado por la derecha en $y=1$) y otro para conteos positivos (>0). $f_{\text{cont}}(y;x,\beta)$ (truncado por la izquierda en $y=1$), que puede ser de tipo Poisson, o Binomial Negativo, (Mullahy, 1986; Cameron, 1998).

$$f_{\text{hurdle}}(y;x,z,\beta,\gamma) = \begin{cases} f_{\text{cero}}(0;z,\gamma) & \text{si } y=0, \\ \frac{(1-f_{\text{cero}}(0;z,\gamma))\cdot f_{\text{cont}}(y;x,\beta)}{(1-f_{\text{cont}}(0;x,\beta))} & \text{si } y>0. \end{cases} \tag{9}$$

Los parámetros del modelo β, γ, y potenciales parámetros de dispersión θ (si f_{cont} o f_{cero} o ambos con densidad negativa binomial) se estiman por Máxima Verosimilitud, donde la especificación de la verosimilitud tiene la ventaja de que los componentes del conteo y de hurdle pueden maximizarse en forma separada.

La regresión sobre la media es:

$$\log(\mu_i) \quad = \quad x_i^\top\beta + \log(1-f_{\text{cero}}(0;z_i,\gamma)) - \log(1-f_{\text{cont}}(0;x_i,\beta)) \tag{10}$$

2.2. Modelos con Exceso de Ceros (MEC)

Los *MEC* (de tipo Poisson (PEC), Binomial Negativa (BNEC)) son modelos de mezcla, que combinan un componente de conteo y una masa de probabilidad en cero, con el restante modelo para los conteos >0 (Cameron, 1998; Hilbe, 2011).

En este caso, hay dos fuentes de cero para el modelo, provenientes de la masa puntual en cero $I_{\{0\}}(y)$ y del modelo de conteo con distribución $f_{\text{cont}}(y;x,\beta)$. La probabilidad de observar un conteo de cero se incrementa con probabilidad $\pi = f_{\text{cero}}(0;z,\gamma)$

$$f_{\text{ceroinfl}}(y;x,z,\beta,\gamma) \quad = \quad f_{\text{cero}}(0;z,\gamma)\cdot I_{\{0\}}(y) \; + \; (1-f_{\text{cero}}(0;z,\gamma))\cdot f_{\text{cont}}(y;x,\beta) \tag{11}$$

donde $I(\cdot)$ es la función indicadora y la probabilidad no observada π de pertenecer al componente de masa puntual se modela con un MLG de tipo binomial $\pi = g^{-1}(z^\top\gamma)$.

La ecuación de regresión para la media es

$$\mu_i \quad = \quad \pi_i\cdot 0 \; + \; (1-\pi_i)\cdot\exp(x_i^\top\beta), \tag{12}$$

usando la función de enlace canónico.

A partir de los diferentes alternativas de modelos planteados hasta aquí se busca el objetivo de identificar los modelos de probabilidad que mejor ajustan al componente C del CPO, de los planteados en la Tabla 1, para luego en contextos de modelos de regresión ver cual es la mejor alternativa, que puede ser un modelo sencillo al trabajar con una sola distribución de probabilidad o modelos más complejos, necesarios para el caso de que exista sobredispersión y exceso de ceros.

Tabla 2: Conjunto de variables regresoras usadas para los modelos de conteo en *RPAFO2015*

Descripción de Variables explicativas		
Variable	Nombre	Descripción
V(1)	CPO	(Nivel de CPO) (C)
V(2)	edad	edad en años(C)
V(3)	sexo	Sexo (2 niveles)
V(4)	niveledu	Nivel educativo (4 niveles)
V(5)	ingresos	Ingresos percibidos (3 niveles)
V(6)	alcohol	Nivel de consumo de alcohol (3 niveles)
V(7)	bebida azucarada	Nro de dias que consume bebidas azucaradas (C)
V(8)	fumaactual	Fuma actualmente (2 niveles)

3. APLICACIÓN DE MODELOS DE CONTEO ALTERNATIVOS PARA LOS COMPONENTES C, P, Y EN RELEVAMIENTO EN POBLACIÓN QUE SE ASISTE FACULTAD DE ODONTOLOGÍA 2015

Para evaluar cómo funcionan las distribuciones y los modelos antes presentados, se trabaja con los datos provenientes del 'Relevamiento en población que se asiste Facultad de Odontología 2015 (RPAFO2015)' estudio sobre personas que demandan atención en la Facultad de Odontología de la Universidad de la República, Uruguay [d], donde se analiza el componente *C*, tratando de identificar su distribución, para luego estimar modelos de regresión usando las siguientes variables explicativas. Este estudio se aplicó a una muestra de 602 personas que consultan en el período que corresponde a mayo 2015-junio 2016, que se seleccionan mediante muestreo sistemático, se les aplica un cuestionario sociodemográfico y un examen completo de la boca, en donde se evalúa el estado de las piezas dentales y de la mucosa, además de medidas antropométricas, de Presión Arterial (PA) y de glicemia (Gli). El tamaño muestral se determinó para poder medir prevalencias de hasta 25 % con un margen de error $\delta = 0.05$ y un nivel de confianza $1 - \alpha = 0.95$ y cubrir hasta una tasa de no respuesta del 90 %. Finalmente, de los 640 originalmente calculados, se obtuvieron 602, que representa una fracción de muestreo de alrededor del 10 % del total de personas que consultan anualmente.

Se detallan las variables explicativas que se usarán para modelar los diferentes componentes del CPO en la Tabla 2.

[d]Pacientes evaluados por los odontólogos del Servicio de registros de la Facultad, desarrollado en el marco del proyecto 'Investigación y Desarrollo' de la Comisión Sectorial de Investigación Científica (CSIC), 2014 de la Universidad de la República

Tabla 3: Medidas de resumen de los componentes de CPO

Componentes	n	$\bar{x}$	sd	mediana	mínimo	máximo
C	602	2.50	3.05	2	0	25
P	602	10.17	8.48	8	0	32
O	602	3.66	4.10	2	0	22
CPO	602	16.33	8.11	17	0	32

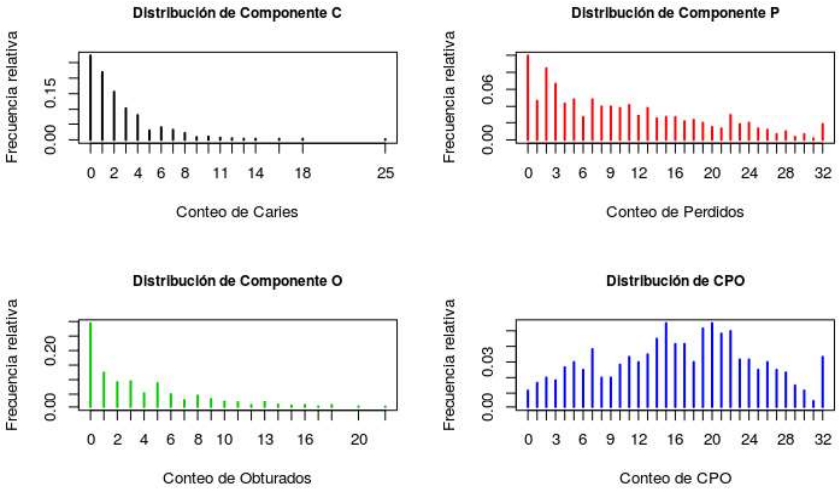

Figura 5: Distribución para el CPO y sus 3 componentes, Fuente: Elaboración propia.

En principio, en la Tabla 3, se tiene la distribución para los tres componentes y el CPO. Si bien se estiman modelos para los tres componentes y el CPO, se muestra con particular detalle lo referente al componente C, dada la forma que presenta en este caso para los datos de la RPAFO2015 y la importancia que desde el punto de vista epidemiológico tiene.

Usando las librerías *COUNT*, (Hilbe, 2016) y *gamlss* (Rigby & Stasinopoulos, 2005) del software R Core Team (2016), se puede estimar la mejor distribución paramétrica para ajustar la variable C. Si bien en los resultados aparecen modelos que consideran inflación de 0, solamente se presentan los cinco modelos paramétricos vistos en la sección 2.

En la Figura 6 se presentan los ajustes hechos para el componente C, considerando que el parámetro μ estimado por la media muestral $\bar{x} = 2.5$, lo que llevaa tener que evaluar una mejor alternativa, siendo que el *PG* es el mejor, tal come se ve en la Tabla 4, donde aparece el criterio de ajuste por el estadístico *AIC* y la Figura 7, que muestra que es la mejor alternativa.

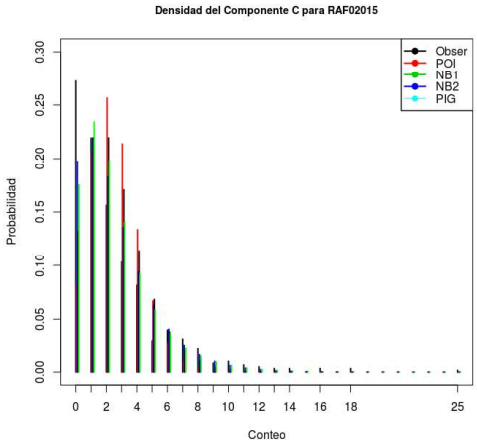

Figura 6: Distribución de diferentes *MCont* para *C*, dado el valor de $\mu = 2.5$, Fuente: Elaboración propia.

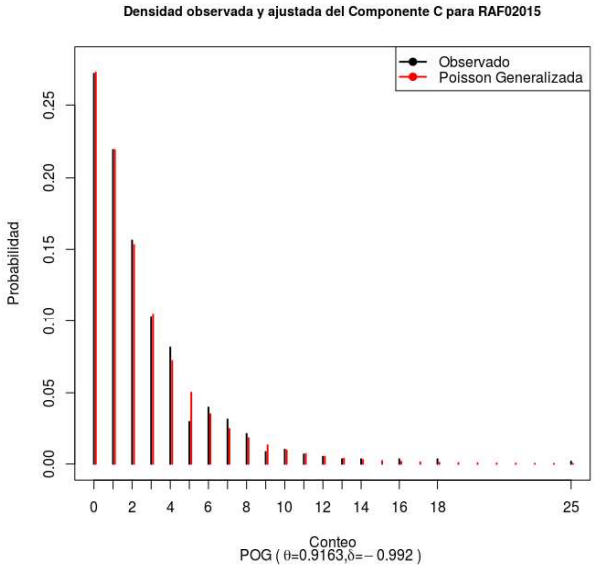

Figura 7: Ajuste del C, dado el valor de $\mu = 2.5$ para un modelo *PG*, Fuente: Elaboración propia.

Tabla 4: Ajuste de la distribución del componente *C*

modelo ajustado para componente C			
tipo de *MCont*	AIC	parámetros	
PG	2522.3	$\mu = 2.5$	$\gamma = 0.370$
NBII	2525.0	$\mu = 2.5$	$\gamma = 2.44$
NBI	2525.1	$\mu = 2.5$	$\gamma = 0.979$
PIG	2525.8	$\mu = 2.5$	$\gamma = 1.23$
PO	3150.2	$\mu = 2.5$	$\gamma = 0$

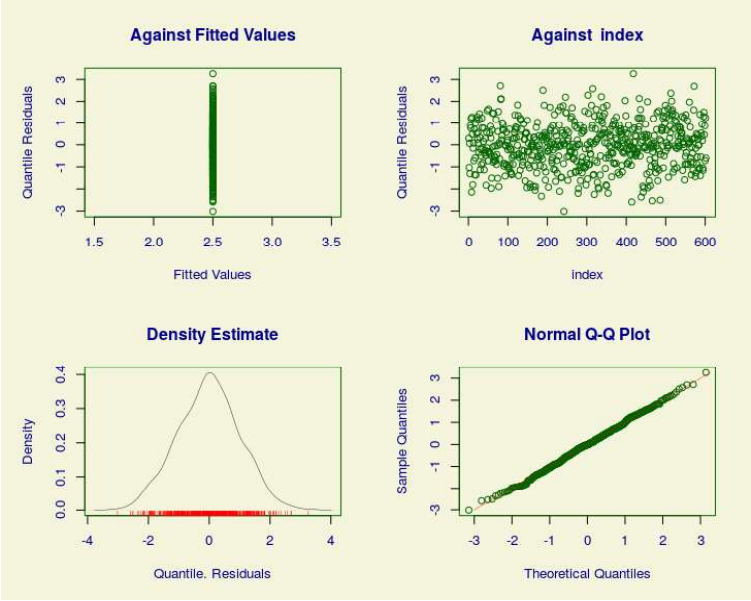

Figura 8: Gráficos de Bondad de ajuste para el modelo de conteo *PG* para C, Fuente: Elaboración propia.

En la Figura 8 puede verse que se cumplen las características para que el ajuste sea adecuado, tal como residuos con media 0, distribución aparentemente gaussiana, a partir de la densidad y cuantiles empíricos que coinciden con los teóricos, donde no aparece un patrón o sesgo en las observaciones. En la Tabla 4 se muestra un resumen de los mejores modelos *MCont* ajustados para el CPO y sus tres componentes. En cada modelo ajustado se realiza el mismo proceso iterativo para encontrar el modelo con mejores indicadores de bondad como el *AIC*, el *SBC* y la devianza global y a su vez se realiza el diagnóstico de ajuste a través de los residuos.

Para entender mejor los resultados encontrados se presenta en la Tabla 5 para cada componente del CPO cual es la bondad de ajuste y la jerarquía que queda para los diferentes *MCont* presentados en sección 2 y donde finalmente, en la Tabla 6, se consigna solamente los *MCont* presentados en detalle previamente. Un aspecto a tener en cuenta es que los coeficientes que se presentan en la Tabla 5 están expresados a través de la función de enlace que es de tipo logarítmico y se usa la base de logaritmos naturales.

Finalmente, antes de pasar a la etapa de elaboración de modelos de pronóstico con las variables regresoras presentadas en la Tabla 2, en la Figura 9 se presentan los mejores modelos ajustados para el CPO y sus 3 componentes.

Tabla 5: Ranking de ajuste de los *MCont* para CPO y sus 3 componentes

Ranking de modelos de conteo por componente	
modelos para componente C	
Modelo	Ranking
Poisson	29
Binomial Negativa (BN - tipo I)	13
Binomial Negativa (BN - tipo II)	12
PIG	14
PG	1
modelos para componente P	
Modelo	Ranking
Poisson	29
Binomial Negativa (BN - tipo I)	9
Binomial Negativa (BN - tipo II)	10
PIG	22
PG	21
modelos para componente O	
Modelo	Ranking
Poisson	29
Binomial Negativa (BN - tipo I)	21
Binomial Negativa (BN - tipo II)	11
PIG	22
PG	21
modelos para componente CPO	
Modelo	Ranking
Poisson	27
Binomial Negativa (BN - tipo I)	10
Binomial Negativa (BN - tipo II)	9
PIG	19
PG	16

Tabla 6: Ajuste de la distribución de CPO y sus 3 componentes

modelo ajustado para componente C				
Tipo	parámetros			
PG	Coeficientes	E.E.	t	Pr($>\|t\|$)
	μ= 0.916	0.049	18.44	< 0.001
	$\gamma = -0.992$	0.0789	-12.56	< 0.001
	Devianza Global =2518.39	AIC=2522.39	SBC=2531.2	
modelo ajustado para componente P				
Tipo	parámetros			
NB-I	Coeficientes	E.E.	t	Pr($>\|t\|$)
	$\mu = 2.32$	0.039	58.63	< 0.001
	$\gamma = -0.168$	0.066	-2.35	0.011
	Devianza Global =4048	AIC=4052	SBC=4060	
modelo ajustado para componente O				
Tipo	parámetros			
NB-II	Coeficientes	E.E.	t	Pr($>\|t\|$)
	$\mu = 1.29$	0.05	$25.16 < 0.001$	
	$\gamma = 1.57$	0.095	16.60	< 0.001
	Devianza Global =2904.1	AIC=2908.1	SBC=2917.1	
modelo ajustado para componente CPO				
Tipo	parámetros			
NB-II	Coeficientes	E.E.	t	Pr($>\|t\|$)
	$\mu = 2.59$	0.023	119.80	< 0.001
	$\gamma = 1.46$	0.078	18.8	< 0.001
	Devianza Global =4304.1	AIC=4308.1	SBC=4316.1	

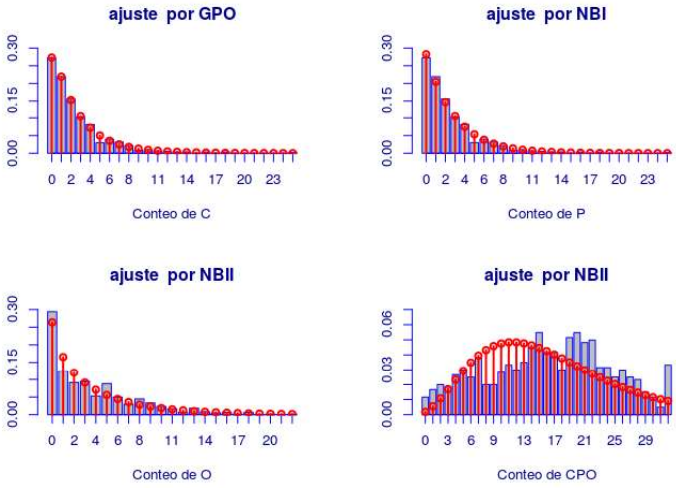

Figura 9: Representación gráfica de los modelos de probablidad ajustados para CPO y sus 3 componentes, Fuente: Elaboración propia.

Tabla 7: Medidas de resumen de las variables regresoras para componente *C*

Medidas de resumen						
Variables	n	$\bar{x}$	sd	mediana	min	max
V(1) CPO	602.00	16.33	8.11	17.00	0.00	32.00
V(2) edad	602.00	45.02	16.85	44.00	18.00	85.00
V(7) bebida azucarada	583.00	3.11	2.95	2.00	0.00	7.00
Variables	Tablas de frecuencia					
V(3) sexo		Femenino :352			Masculino:250	
V(4) niveledu	1: 170	2: 175	3: 162	4: 93	Sin datos: 20	
V(5) ingresos	1: 325	2: 189	3: 58	Sin datos: 30		
V(6) alcohol	No consume: 214	mensual: 364	semanal/diario: 21	Sin datos: 3		
V(8) fumactual	No : 399	SI: 199	Sin datos: 4			

En la Tabla 7 se presentan las medidas de resumen para las variables regresoras del componente C, ya que solamente por ser de los tres el más relevante desde el punto de vista epidemiológico y el más frecuentemente estudiado, será el único analizado mediante modelos de regresión en este trabajo.

El modelo estimado presentado en la Tabla 8 toma en cuenta la sobredispersión que existe para el conteo de Caries, que este caso es de casi 2.5. Observando el modelo se podría pensar que existe una relación entre el número de *C* y el sexo, la edad, la ingesta de bebidas azucaradas y las personas que tienen un mayor ingreso.

Tabla 8: Modelo de regresión quasi-Poisson para componente *C*

	Coeficiente	EE	valor z	Pr(> |z|)
(Intercepto)	-0.349	0.153	-2.27	0.0234
CPO	0.0578	0.007	8.00	0.0000
cedad	-0.034	0.004	-8.62	0.0000
sexo=Masculino	0.285	0.085	3.33	0.0009
bebida azucarada	0.049	0.015	3.23	0.0013
ingresos(2)	-0.127	0.094	-1.36	0.175
ingresos(3)	-0.378	0.170	-2.22	0.026

Parámetro de Dispersión =2.463

Tabla 9: Modelo de regresión NBI para componente C

	Coeficiente	EE	valor z	Pr(> |z|)
(Intercepto)	-0.4259	0.1526	-2.79	0.0052
CPO	0.0621	0.0076	8.17	0.0000
edad	-0.0322	0.0039	-8.28	0.0000
sexo=Masculino	0.3126	0.0899	3.48	0.0005
bebida azucarada	0.0461	0.0156	2.95	0.0032
ingresos(2)	-0.0928	0.0966	-0.96	0.3367
ingresos(3)	-0.3956	0.1616	-2.45	0.0143

Parámetro de Dispersión para quasipoisson =2.463

Tabla 10: Modelo de regresión GPO para componente C

Coeficientes para parámetro θ				
Variables	Coeficiente	EE	valor z	Pr(> |z|)
(Intercepto)	-0.455	0.162	-2.804	0.0052
CPO	0.0637	0.008	7.487	< 0.0001
edad	-0.0321	0.0038	-8.237	< 0.001
sexo=Masculino	0.319	0.091	3.496	< 0.001
bebida azucarada	0.0455	0.016	2.833	< 0.005
ingresos(2)	-0.088	0.097	-0.909	0.363
ingresos(3)	-0.396	0.158	-2.505	0.012
Coeficientes para parámetro δ				
Variables	Coeficientes	EE	valor z	Pr(> |z|)
(Intercepto)	-1.472	0.105	-14.02	< 0.001

Trabajando con un modelo de probabilidad de Tipo NBII en función de los valores consignados en la columna Coeficiente, Error Estándar (EE), (que ya se vio que ajusta mejor a los datos) como aparecen en la Tabla 5, los resultados aparecen en la Tabla 9.

Si en lugar de estimar un modelo *MH* con distribución Poisson para el componente truncado se usa la distribución *BN*, donde aparece el parámetro de dispersión, los resultados cambian tal como se muestra en la Tabla 12 y donde ambos se pueden comparar para observar la mejoría en usar el modelo más complejo a través del test de Wald, que muestra que los cambios al usar el modelo con BN para el componente truncado no son relevantes.

Tabla 11: Modelo de regresión *MH* para componente C, con distribución Poisson

Modelo de Conteo *MH* con distribución Poisson				
Coeficientes para modelo con función de enlace logit				
Variables	Coeficiente	EE	valor z	Pr(> |z|)
(Intercepto)	1.715	0.316	5.42	< 0.0001
CPO	0.057	0.017	3.33	< 0.0001
edad	-0.035	0.007	-4.50	< 0.0001
sexo=Masculino	0.438	0.209	2.09	0.0363
ingresos(2)	-0.139	0.220	-0.63	0.527
ingresos(3)	-0.835	0.310	-2.68	0.007
Modelo de Conteo con distribución Poisson truncado				
Coeficientes del modelo				
Variables	Coeficiente	EE	valor z	Pr(> |z|)
(Intercepto)	1.27	0.112	11.25	< 0.0001
CPO	0.057	0.005	11.13	< 0.0001
edad	-0.030	0.002	-11.18	< 0.0001
sexo=Masculino	0.244	0.059	3.82	< 0.0001
bebida azucarada	0.045	0.010	4.27	< 0.0001

Tabla 12: Modelo de regresión *MH* para componente C, con distribución Binomial Negativa

Modelo de Conteo *MH* con distribución *BN*				
Coeficientes para modelo con función de enlace logit				
Variables	Coeficiente	EE	valor z	Pr(> |z|)
(Intercepto)	1.715	0.316	5.42	< 0.001
CPO	0.057	0.017	3.33	< 0.001
edad	-0.035	0.007	-4.50	< 0.001
sexo=Masculino	0.438	0.209	2.09	0.036
ingresos(2)	-0.139	0.220	-0.63	0.527
ingresos(3)	-0.835	0.310	-2.68	0.007
Coeficiente $\sigma = 2.034$ para modelo *BN*				
Modelo de Conteo con distribución *BN* truncada				
Coeficientes del modelo				
Variables	Coeficiente	EE	valor z	Pr(> |z|)
(Intercepto)	0.968	0.188	5.12	< 0.001
CPO	0.071	0.009	7.31	< 0.001
edad	-0.034	0.004	-7.61	< 0.001
sexo=Masculino	0.270	0.100	2.69	< 0.001
bebida azucarada	0.046	0.018	2.55	0.011
Log(σ)	0.712	0.208	3.41	< 0.001

Finalmente en la Tabla 13 puede verse un resumen con la performance de los diferentes modelos, donde se presentan además indicadores de ajuste, donde la mejor opción es el modelo con obstáculos y distribución de conteo truncado de tipo *BN*, aunque igual subestima la cantidad de personas sanas de Caries, lo que estaría marcando que el modelo hallado parece adecuarse a una población más enferma, por lo que se plantea la necesidad de seguir indagando para identificar un mejor modelo.

Tabla 13: Performance de los diferentes modelos de regresión para componente C, usando modelos paramétricos (MLG) y modelos con obstáculos

	Modelos vía (*MLG*)		Modelos con obstáculos	
Tipo	Poisson	BN	Poisson *MH*-Hurdle	BN-*MH*
# de parámetros	7	8	13	12
AIC	2471	2182	2340	2212
SBC	2500	2216		
$Log(\mathcal{L})$	-1228	-1083	-1156	-1094
$\sum_1^{602} x_i = 0$	82	143	146	146

4. DISCUSIÓN SOBRE LA DISTRIBUCIÓN Y EL MODELADO DE LOS COMPONENTES DE CPO

Luego de presentados los diferentes modelos de probabilidad en la sección 2, puede verse la forma de la distribución de los 4 componentes del estudio *RPAFO2015*, donde en primer lugar en esta sección se discute sobre modelado del componente C dada la importancia de éste desde el punto de vista epidemiológico y de la Salud Pública, para luego presentar el comportamiento observado para los restantes componentes del CPO.

Puede verse por lo tanto que si se trabaja solamente con la distribución de C con independencia del resto de las variables (es decir la distribución incondicional o en un contexto sin variables regresoras), puede decirse que el mejor *MCont* es el que corresponde a la *BN*-II, como se ve gráficamente en la Figura 6, pero sin perder de vista que para los 5 modelos paramétricos presentados en 2, el ajuste es pobre.

Por tales motivos tal como se adelantó en la sección 3, la mejor alternativa es la *PG*, tal como se presenta en la Tabla 6, donde de un total de 29 distribuciones de conteo que se pueden estimar a través de la librería *gamlss*, (Rigby & Stasinopoulos, 2005), la *PG* está en el primer lugar. Entre todas esas distribuciones de conteo están las básicas presentadas en la sección 2, que corresponden a las ecuaciones 2.1 y 2.2, y varias variantes de las mismas que dada su complejidad matemática se dejan de lado. En la Figura 9, se presenta la calidad del ajuste que muestra que los residuos tienen media 0, distribución aparentemente gaussiana, a partir de la densidad y cuantiles empíricos que coinciden con los teóricos, donde no aparece un patrón o sesgo en las observaciones.

Si ahora se considera el resto de los componentes del CPO, el ajuste de éstos a través de *MCont* básicos es insuficiente, salvo para el componente P. Para el caso de O, con un ajuste por una *BN*-II los datos, muestran un exceso de 0 y una diferencia importante. Para terminar esta parte del análisis el CPO es la variable de conteo que dado su comportamiento de no monotonía y ser multimodal, parece difícil de responder a un modelo con distribución conocido, sino que parece adecuado considerar a un proceso de mezcla.

En cambio en un contexto de regresión en las Tablas 9 y 10 se presentan los resultados de ajustar por un *MPoi* y un *BN*, dada la sobredispersión que existe y donde de las 8 variables regresoras previamente consideradas en la Tabla 3, solamente resultan significativas el CPO, la edad, el sexo, la cantidad de días en la semana donde se ingiere bebidas azucaradas y el ingresos de los individuos participantes del estudio. Para ambos modelos, donde los coeficientes muestran valores similares, se consideran las variables que son significativas y la asociación con el logaritmo del número medio de caries muestra resultados esperables. La cantidad de Caries se asocia con un mayor nivel de CPO, con un aumento promedio de una Caries por cada punto extra en el CPO, con un decremento de casi una Caries por cada año por encima de la media, lo que podría explicarse porque al aumentar la edad las personas tiene menos piezas presentes; a su vez el consumo diario de bebidas azucaradas tiene un aumento promedio de una caries por cada día en que la persona consume bebidas. Un aspecto importante para ambos modelos es la importancia del coeficiente asociado a sexo masculino, donde el número medio de Caries es de casi 1.36 con respecto a las mujeres. Finalmente los ingresos muestra una asociación negativa donde las personas que están en el último tramo de ingresos tienen una reducción en el número de Caries importante con respecto a los del primer tramo, que es la referencia.

Todos estos resultados son muy similares para ambas versiones de los modelos y podrían resultar en la elaboración de teoría epidemiológica que para el investigador en Biomedicina no advertido podría llevarlo a cometer errores, si no se toma en cuenta un aspecto que en general no se considera y es la capacidad predictiva del modelo estimado.

Para este caso donde el conteo muestra comportamiento patológico al tener sobredispersión y exceso importante de 0, es fundamental verificar el grado de ajuste y no alcanza por lo tanto que las variables sean significativas, ya que no tomar en cuenta este aspecto puede llevar al investigador en Biomedicina que trabaje con modelos similares a cometer errores muy importantes pautando asociaciones entre variables que en la práctica no se dan.

En particular interesa ver que sucede con el conteo de 0. El componente C muestra que hay 164 personas sanas de Caries, es decir con un conteo=0, mientras que los 2 modelos básicos estimados para el caso de Poisson pronostican 82 personas sin Caries y el modelo de la *BN* el conteo es de 143. Por eso motivo los modelos con obstáculos o *MH* con distribución de conteo truncada que se presentan en las Tablas 11 y 12, muestran una mejor capacidad de detectar personas con C=0 y a su vez las variables regresoras de cada parte del modelo pautan un aspecto muy importante que se detalla a continuación.

La parte del modelo MH que se modela mediante un logit y es el que permite saltar el obstáculo, es el muestra que perfil tiene la persona para tener o no Caries, que en el caso del MH Poisson son el CPO, la edad, el sexo, el nivel de ingesta de bebidas azucaradas y el ingreso, mientras que para la parte del modelo truncado la ingesta no se debe tener en cuenta y el ingreso pasa a tener un efecto contrario al cambiar de signo.

Para el caso del modelo MH Binomial Negativo, las variables que permiten saltar el obstáculo a través del modelo logit son CPO, edad, sexo e ingreso, mientras que para la parte del modelo truncado desaparece el ingreso, mientras que la ingesta de bebidas azucaradas aparece como una variable moduladora en el aumento del conteo de Caries.

5. CONCLUSIONES PARA LOS *MCONT* PARA EL ESTUDIO RPA-FO2015

De los resultados encontrados resulta fundamental recordar los siguientes pasos que debería seguir el investigador biomédico al trabajar con este tipo de datos que presentan varias patologías desde el punto de vista estadístico, por lo cual no se pueden usar los modelos básicos de conteo.

- Previamente examinar cuáles pueden ser las distribuciones que reproducen los conteos (en este caso C,P,O) independientemente de los modelos que se deseen elaborar para encontrar asociaciones;
- No alcanza con encontrar variables significativas con las cuales desarrollar teoría epidemiológica que tenga sentido para el investigador, ya que estaría basada en modelos que no son comparables con los datos bajo estudio; en el caso presentado las asociaciones son válidas para poblaciones menos enfermas (hay menos gente con Caries);
- Es necesario usar modelos combinados más complejos como los que se componen de 2 submodelos
    - Uno que trabaja con personas que no tienen Caries, aspecto que se modela con el componente 1 del modelo (MH);
    - Cuando tienen Caries, la cantidad de éstas se modelan con el componente 2 del modelo (MH).

Podría suceder que las variables regresoras que se usan para salir del obstáculo no necesariamente sean las mismas que las que contribuyen a modelar el conteo truncado (> 0) e incluso siendo las mismas pueden cambiar el sentido de la asociación o la intensidad de las mismas con coeficientes con distintos valores.

Si bien no se trabajó con el resto de los componentes del CPO o el CPO mismo en un contexto de regresión ya se vió que las distribuciones no son identificables fácilmente a través de un modelo explícito de los presentados, sino que hay que pensar en identificar mezclas de distribuciones.

Como último comentario siempre es deseable manejar modelos parsimoniosos pero es deber del investigador conocer las limitaciones de los mismos y lograr un equilibrio entre modelo sencillo pero adecuado.

Referencias

Breilh, J. (2010). La epidemiología crítica: una nueva forma de mirar la salud en el espacio urbano. *Salud Colectiva*, 31, 152-157. [Consultada 2021]. Disponible en: https://doi.org/10.18294/sc.2010.359.

Cameron, A. & Trivedi, P. (1998). Regression analysis of count data. Cambridge, UK New York, NY, USA: Cambridge University Press, ISBN: 0521635675 paperback, 432 pages.

Fejerscov, O. (2004). Changing paradigms in concept son dental Caries: Consequences for oral health care. *Caries Research*, 38, 182-191, [Consultada 2021]. Disponible en: https://doi.org/10.1159/000077753.

Giner, G. y Smyth, G. K. (2016). statmod: probability calculations for the inverse gaussian distribution. *R Journal*, 8(**1**), 339–351.

Hilbe, J. (2011). Negative binomial regression. Cambridge University Press.

Hilbe, J. M. (2016). COUNT: Functions, Data and Code for Count Data. R package version 1.3.4.

Holst, D., Schuller, A., Aleksejunienė, J., & EriKsen, H. (2001). Caries in populations–a theoretical, causal approach. European journal of oral sciences, 109(**3**), 143–48.[Consultada 2021]. Disponible en: https://doi.org/10.1034/j.1600-0722.2001.00022.x

Maltz, M., Jardim, J. J., & Alves, L. S. (2010). Health promotion and dental Caries. Brazilian oral research, 24 Suppl 1, 18–5. [Consultada 2021]. Disponible en: https://doi.org/10.1590/s1806-83242010000500004

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics, 33(**3**), 341–65.

Petersen P. E. (2004). Challenges to improvement of oral health in the 21st century–the approach of the WHO Global Oral Health Programme. International dental journal, 54(**6 Suppl 1**), 329–343. [Consultada 2021]. Disponible en: https://doi.org/10.1111/j.1875-595x.2004.tb00009.x

R Core Team (2016). A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rigby, R., & Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(**3**), 507-554. [Consultada 2021]. Disponible en: https://doi.org/10.1111/j.1467-9876.2005.00510.x

Wheeler, B. (2016). SuppDists: Supplementary Distributions. R package version 1.1-9.4.

Whitmore, A. G. (1975). The inverse Gaussian distribution as a model of hospital stay. Health Services Research,10(**3**), 297–302.

Whitmore, A. G. & Yalovsky, M.(978). A normalizing logarithmic transformation for inverse Gaussian random variables. Technometrics, 20(**2**), 207–208, [Consultada 2021]. Disponible en: https://www.tandfonline.com/doi/abs/10.1080/00401706.1978.10489648.