El Bootstrap paramétrico y no parametrico y su aplicación en los modelos log-lineal Poisson

Antonio Bravo Quiroz [*]

Universidad Ricardo Palma

23 de noviembre de 2017

Índice

[*]abravoqz@gmail.com

Resúmen:

Los modelos lineales generalizados son una clase de técnicas estadísticas para el análisis de la relación funcional entre uno o más variables independientes o variables regresoras, con una variable dependiente o respuesta, y unifica en una sola clase los modelos lineales con errores normales y no normales, todas ellas perteneciente a la familia exponencial a un parámetro.

Los modelos log-lineales constituyen una técnica estadística, integrante de la clase de los modelos lineales generalizados, que permite el análisis de los datos de una tabla de contingencia, en la búsqueda de la asociación entre los factores o clases de dos o más variables categóricas, sin distinguir si son variables regresoras o respuestas, donde el análisis es equivalente al ANOVA para la variable respuesta con errores normales.

Considerando que la distribución multinomial o la producto multinomial es la distribución natural para el análisis de una tabla de contingencia, la misma que depende de las contadas n_{ij} y de las probabilidades de clasificación π_{ij}, estos componentes o parámetros de la distribución multinomial los podemos asociar con la distribución Poisson con media μ_{ij}, que es una distribución asociada con el número de ocurrencias de un evento $y_{ij} = n_{ij}$ en una unidad de tiempo o espacio de observación, las mismas que ocurren con una probabilidad $\pi_{ij} = P(Y_{ij} = y_{ij})$. Así, las contadas n_{ij} de las celdas de una tabla de contingencia, que asumen valores enteros no negativos, las podemos asociar con la distribución Poisson, bien como una aproximación de una binomial con la distribución Poisson o que las contadas ocurren como una

realización de un proceso de Poisson con espacio de observación $(0, t]$.

Además, la inclusión de la distribución Poisson en el análisis log-lineal facilita el análisis de los datos, dado que la expresión del predictor lineal $\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$, tiene las características de de una regresión lineal múltiple cunado las variables regresoras son continuas y tienen la forma de los modelos ANOVA si las variables regresoras son categóricas.

La estimación bootstrap en los modelos log-lineal Poisson, permiten mejorar las estimaciones del error estándar de un estimador $\hat{\theta}$, básicamente cuando la data es insuficiente, pero consistente, permitiendo mejorar las estimaciones de los intervalos de confianza y de la normalidad asintótica de los estimadores de máxima verosimilitud. En el presente trabajo de investigación lo hemos usado para mejorar la estimación del p-valor de las estadísticas de test chi-cuadrado de Pearson y del test de la razón de máxima verosimilitud, que en el caso de tablas de contingencia de dos vías, asintóticamente tienen distribución chi-cuadrado con $(R-1)(S-1)gl$.

Asimismo, la contribución del trabajo de investigación, es proveer a los interesados de un material de lectura teórica para la difusión, entendimiento y uso de los modelos lineales generalizados, de los modelos log-lineal y de la técnica del bootstrap, considerando que su aplicación es interesante, para lo cual se requieren bases de datos adecuada que no permitan celdas con contadas menores a cinco o nulas.

Para la aplicación de la teoría estudiada se buscó y probó diferentes bases de datos, entre ellos, datos de la encuesta de hogares del INEI, como modelo de aplicación de las técnicas estudiadas daban resultados irrelevantes; pero, con suerte nos encontramos con el trabajo De los Rios y Bravo (2012) de manera casual, quienes cedieron la base de datos original del trabajo de investigación sobre secuelas de la tuberculosis. El análisis de dicha data, de por sí, el tema es facinante, pero nos enfrentamos a una realidad donde el tamaño de la muestra era pequeña, por lo que se trabajó arduamente, con la técnica de recolección de datos observacionales retrospectivos para el acopio de nuevos datos, trabajo que requirió una dedicación de largo aliento, pudiendo mejorar el tamaño de la muestra a 85 observaciones.

Palabras clave:

modelo lineal generalizado, modelos log-lineal, aproximación de un estimador por el método de Newton-Rapson, estimación Bootstrap, secuela de TBC.

1. Conceptos Preliminares

1.1. Tablas de contingencia

Una tabla de contingencia de dos vías permite el estudio de la asociación o influencia entre dos características A y B observables en cada una de las unidades de la población, que las podemos asociar con las variables categóricas (X_A, X_B), donde una es la variable independiente X_A y a otra es la variable dependiente o respuesta X_B, con probabilidades de clasificación π_{ij}, y contadas m_{ij}, para $i = 1, 2, \cdots, R$ y $j = 1, 2, \cdots, S$. Esto es, las probabilidades de clasificación de un individuo de la población en cada una de las celdas de la tabla de contingencia, son:

Categorías del Factor X_A:	Categorías del Factor X_B: X_{B1}	X_{B2}	$\cdots$	X_{BS}	Total
X_{A1}	π_{11}	π_{12}	$\cdots$	π_{1S}	π_{1+}
X_{A2}	π_{21}	π_{22}	$\cdots$	π_{2S}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_{AR}	π_{R1}	π_{R2}	$\cdots$	π_{RS}	π_{R+}
Total	π_{+1}	π_{+2}	$\cdots$	π_{+S}	1

Las unidades de la población clasificadas en la tabla de contingencia nos da como resultado las contadas m_{ij}, que las presentamos en la siguiente tabla de contingencia:

Categorías del Factor X_A:	Categorías del Factor X_B: X_{B1}	X_{B2}	$\cdots$	X_{BS}	Total
X_{A1}	m_{11}	m_{12}	$\cdots$	m_{1S}	m_{1+}
X_{A2}	m_{21}	m_{22}	$\cdots$	m_{2S}	m_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_{AR}	m_{R1}	m_{R2}	$\cdots$	m_{RS}	m_{R+}
Total	m_{+1}	m_{+2}	$\cdots$	m_{+S}	m_{++}

Dado que la población no es conocida, no es posible conocer las probabilidades π_{ij}, ni las contadas m_{ij}. Luego, el valor de los parámetros π_{ij} y m_{ij} los aproximaremos por medio de sus estimadores, a través de una muestra aleatoria. Para esto, consideremos una muestra de unidades d ve la población de tamaño n, clasificados en una tabla de contingencia de dos vías, según los factores de clasificación X_A e X_B:

Categorías del Factor X_A	Categorías del Factor X_B: X_{B1}	X_{B2}	$\cdots$	X_{BS}	Total
X_{A1}	n_{11}	n_{12}	$\cdots$	n_{1S}	n_{1+}
X_{A2}	n_{21}	n_{22}	$\cdots$	n_{2S}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_{AR}	n_{R1}	n_{R2}	$\cdots$	n_{RS}	n_{R+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+S}	n_{++}

donde:

n_{ij} son las contadas observadas en la categoría i del factor independiente X_A y la categoría j del factor respuesta X_B. Esto es,

$$n_{ij} = \#\left\{u_k \in (X_{Ai}, X_{BJ}); k = 1, \cdots, n\right\} \qquad (1.1)$$

para $i = 1, \cdots, R;\ j = 1, \cdots, S$.

n_{i+} y n_{+j} son las contadas marginales fila y columna, respectivamente, tal que

$$n_{i+} = \sum_{j=1}^{S} n_{ij} \qquad \text{y} \qquad n_{+j} = \sum_{i=1}^{R} n_{ij}$$

n_{++} son las contadas totales, tal que

$$n_{++} = \sum_{i=1}^{R} \sum_{j=1}^{S} n_{ij} = n$$

1.2. Modelos de muestreo multinomial

Bajo el muestreo multinomial, se pueden presentar dos modelos que acondicionan la muestra en una tabla de contingencia:

- El *modelo de muestreo multinomial completo*, donde las celdas (i, j) de la tabla de contingencia las asociamos con una única distribución multinomial, donde el total $n_{++} = n$ es fijo y conocido, y n es el tamaño de la muestra.
- El *muestreo producto multinomial*, donde cada una de las filas de la tabla de contingencia están asociadas con grupos de clasificación independientes y constituyen una multinomial con total de las filas n_{i+} fijas y conocidas.

1.2.1. La distribución multinomial completa

Si las n unidades de la muestra son clasificadas en una tabla de contingencia de dos vías, de dimensión $R \times S$ de acuerdo a las características X_A y X_B, la distribución conjunta de las $R \times S$ contadas n_{ij} tienen distribución multinomial, con función de probabilidad

$$P\left((\mathbf{X_A}, \mathbf{X_B}) = \mathbf{n}\right) = \frac{n_{++}!}{\prod_{ij} n_{ij}!} \prod_{i=1}^{R} \prod_{j=1}^{S} \pi_{ij}^{n_{ij}} \tag{1.2}$$

con $\mathbf{n} = (n_{11}, n_{12}, \cdots, n_{RS})$, y

$$n_{ij} \geq 0 \quad \text{y} \quad \sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij} = n_{++}, \text{ fijo;}$$
$$\pi_{ij} \geq 0 \quad \text{y} \quad \sum_{i=1}^{R}\sum_{j=1}^{S} \pi_{ij} = 1$$

En este caso las contadas marginales fila n_{I+} y las contadas marginales columna n_{+j} son aleatorias y el total general $n_{++} = n$ fijo.

1.2.2. La distribución producto multinomial

Sea X_A una variable categórica de exposición con R categorías o grupos independientes, y X_B es la variable categórica respuesta al factor de exposición, con S categorías. Si la muestra de n unidades es estratificado en R grupos de tamaño n_{i+} fijo, para $i = 1, 2, \cdots, R$, entonces, las contadas $n_{i1}, n_{i2}, \cdots, n_{iS}$ del grupo i tienen distribución conjunta multinomial y, por la independencia de los grupos de exposición, la distribución conjunta de los R grupos o categorías de la variable exposición X_A tienen distribución producto multinomial, con función de probabilidad conjunta.

$$\begin{aligned} P\left((\mathbf{X_A}, \mathbf{X_B}) = \mathbf{n}\right) &= \prod_{i=1}^{R} P\left((\mathbf{X_{Ai}}, \mathbf{X_{Bj}}) = \mathbf{n}_i\right) \\ &= \prod_{i=1}^{R} \left\{ \frac{n_{i+}!}{n_{i1}! \cdots n_{iS}!} \prod_{j=1}^{S} \pi_{ij}^{n_{ij}} \right\} \end{aligned} \tag{1.3}$$

donde $\mathbf{n}_i = (n_{i1}, n_{i2}, \cdots, n_{iS})$ y para $i = 1, 2, \cdots, R$,

$$n_{ij} \geq 0 \quad \text{y} \quad \sum_{j=1}^{S} n_{ij} = n_{i+}, \text{ fijo;}$$
$$\pi_{ij} \geq 0 \quad \text{y} \quad \sum_{j=1}^{S} \pi_{ij} = 1$$

1.3. Estimación y pruebas de hipótesis

Si la muestra es observada y clasificada en la tabla de contingencia, la función de probabilidad de la multinomial definido en (1.2) y en (1.3) se transforma en la función de verosimilitud de la muestra, con contadas n_{ij} conocidas y fijas, que es

$$\mathcal{L}(\boldsymbol{\pi}) = \prod_{i=1}^{R} \left\{ \frac{n_{i+}!}{n_{i1}! \cdots n_{iS}!} \prod_{j=1}^{S} \pi_{ij}^{n_{ij}} \right\}$$

y por el principio de la verosimilitud fuerte, Bickel & Doksum [1] y Cox & Hinklery [3], tanto la función de verosimilitud y su logaritmo alcanzan un máximo en el mismo punto, que en la práctica posibilita obtener el estimador con mayor facilidad. Luego, el logaritmo de la función de verosimilitud, sin considerar los términos que no son función de las probabilidades π_{ij}, es,

$$\mathbf{L}(\boldsymbol{\pi}) = \sum_{i=1}^{R} \sum_{j=1}^{S} n_{ij} \log \pi_{ij} \tag{1.4}$$

1.3.1. Estimación bajo la distribución multinomial

Bajo la distribución multinomial, el máximo de (1.4) se alcanza en el punto $\hat{\boldsymbol{\pi}}$, donde

$$\hat{\boldsymbol{\pi}} = (\hat{\pi}_{11}, \hat{\pi}_{12}, \cdots, \hat{\pi}_{RS}) = \left(\frac{n_{11}}{n_{++}}, \frac{n_{12}}{n_{++}}, \cdots, \frac{n_{RS}}{n_{++}} \right)$$

así, $\hat{\boldsymbol{\pi}}$ es el estimador de máxima verosimilitud para las probabilidades de clasificación $\boldsymbol{\pi} = (\pi_{11}, \pi_{21}, \cdots, \pi_{RS})$, que por ser un estimador de máxima verosimilitud satisfacen las propiedades de consistencia (Silvey [4, pág. 76]), en el sentido que, como las proporciones $\hat{\pi}_{ij} = n_{ij}/n$ tienen distribución binomial, convergen fuertemente a π_{ij}, cuando $n \to \infty$.

Bajo el principio de la invarianza de los estimadores de máxima verosimilitud (Bickel

& Doksum [1]) y (Cox & Hinkley [3]), el estimador de máxima verosimilitud de las contadas m_{ij}, son

$$\hat{m}_{ij} = n_{++} \cdot \hat{\pi}_{ij}$$

Considerando el modelo irrestricto, el estimador de máxima verosimilitud para las contadas m_{ij}, son

$$\hat{m}_{ij} = n_{++} \cdot \hat{\pi}_{ij} = n_{++} \cdot \frac{n_{ij}}{n_{++}} = n_{ij} \tag{1.5}$$

Si imponemos la restricción que los factores de clasificación X_A e X_B son independientes, implica el contraste de la hipótesis nula

$$H_0 : \ \pi_{ij} = \pi_{i+} \cdot \pi_{+j};; \quad i = 1, 2, \cdots, R, \quad j = 1, 2, \cdots, S$$

Luego, bajo la hipótesis nula H_0, los estimadores de máxima verosimilitud para las probabilidades π_{ij} y las contadas teóricas m_{ij}, son, respectivamente,

$$\hat{\pi}_{ij} = \hat{\pi}_{i+} \cdot \hat{\pi}_{+j} = \frac{n_{i+}}{n_{++}} \cdot \frac{n_{+j}}{n_{++}}$$

y

$$\hat{m}_{ij} = n_{++} \cdot \hat{\pi}_{ij} = \frac{n_{i+} \cdot n_{+j}}{n_{++}} \tag{1.6}$$

1.3.2. Estimación bajo la distribución producto multinomial

Si la muestra correspondientes a los R grupos independientes han sido observadas, el logaritmo de la función de verosimilitud de la muestra, ignorando los términos que no dependen del parámetro π_{ij}, está dado por

$$\mathbf{L}(\boldsymbol{\pi}) = \sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij}\log\pi_{ij} \tag{1.7}$$

que es equivalente a lo obtenido para el modelo multinomial (1.4). Si $\hat{\boldsymbol{\pi}}_i$ maximiza la función de verosimilitud $L(\boldsymbol{\pi})$, donde

$$\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}, \hat{\pi}_{i2}, \cdots, \hat{\pi}_{iS}) = \left(\frac{n_{i1}}{n_{i+}}, \frac{n_{i2}}{n_{i+}}, \cdots, \frac{n_{iS}}{n_{i+}}\right)$$

para $i = 1, 2, \cdots, R$. Entonces, $\hat{\boldsymbol{\pi}}_i$ es el estimador de máxima verosimilitud para $\boldsymbol{\pi}_i$. Por el principio de la invarianza de los estimadores de máxima verosimilitud, el estimador de máxima verosimilitud para las contadas m_{ij}, son

$$\hat{m}_{ij} = n_{i+}\cdot\hat{\pi}_{ij} = n_{ij}; \quad i = 1, 2, \cdots, R$$

Ahora, modificamos el modelo irrestricto con la hipótesis de asociación de factores:

$$H_0: \ \pi_{1j} = \pi_{2j} = \cdots = \pi_{Rj}; \quad \text{para} \quad j = 1, 2, \cdots, S$$

Si la hipótesis H_0 es verdadera, implica que podemos obtener un estimador común π_j para cada una de las columnas de la tabla, donde $\pi_j = \pi_{1j} = \pi_{2j} = \cdots = \pi_{Rj}$, de modo que el logaritmo de la función de verosimilitud es

$$\mathbf{L}(\boldsymbol{\pi}) = \sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij}\log\pi_j = \sum_{j=1}^{S} n_{+j}\log\pi_{ij}$$

En este caso, el estimador de máxima verosimilitud para $\boldsymbol{\pi}$, bajo hipótesis nula de asociación de factores, es

$$\hat{\pi}_{ij} = \hat{\pi}_j = \frac{n_{+j}}{n_{++}}$$

y, nuevamente, por la propiedad de invarianza de los estimadores de máxima verosimilitud, bajo la hipótesis nula H_0 el estimador para $m_{ij} = n_{i+}\pi_{ij}$, es

$$\hat{m}_{ij} = n_{i+} \cdot \hat{\pi}_{ij} = n_{i+} \frac{n_{+j}}{n_{++}} = \frac{n_{i+} \cdot n_{+j}}{n_{++}} \tag{1.8}$$

Podemos observar que los resultados (1.6) y (1.8) son equivalentes, esto implica tanto la hipótesis de independencia y la hipótesis de asociación los podemos contrastar de la misma manera, en el sentido que rechazar la hipótesis de independencia, implica aceptar la hipótesis de asociación, y viceversa.

Luego, la estadística de test natural para contrastar la hipótesis nula es la estadística chi-cuadrada de Pearson, que mide las discrepancias entre las contadas observadas en la muestra y las contadas esperadas bajo la hipótesis nula H_0, que es

$$X_0^2 = \sum_{i=1}^{R} \sum_{j=1}^{S} \frac{\left(n_{ij} - \hat{m}_{ij}^0\right)}{\hat{m}_{ij}^0} \ \sim \ \chi^2_{(R-1)(S-1)gl} \tag{1.9}$$

Una estadística de test alternativa para probar la hipótesis nula H_0es el test de razón de verosimilitud o deviance, que también mide la diferencia entre las contadas observadas y las esperadas obtenidas bajo el modelo no restringido y bajo el modelo restringido por la hipótesis nula, que está dado por la expresión

$$D = 2 \sum_{i=1}^{R} \sum_{j=1}^{S} n_{ij} \log \left(\frac{n_{ij}}{\hat{m}_{ij}} \right) \ \sim \ \chi^2_{(R-1)(S-1)gl} \tag{1.10}$$

1.4. La distribución de muestreo Poisson

La distribución Poisson proporciona la distribución del número de ocurrencias de un evento de interés, observados en una unidad de tiempo o espacio fijo, donde cada realización es independiente de los demás y con una probabilidad de ocurrencia pequeña $\pi \to 0$.

Una característica de la distribución Poisson es que depende de un único parámetro λ, que indica la razón de las ocurrencias por unidad de tiempo o espacio y, por

la probabilidad de ocurrencia que es pequeña, está asociado con la realización de eventos raros.

Una variable aleatoria Y que asume valores enteros no negativos $0, 1, 2, \cdots, \infty$ tiene distribución Poisson con parámetro λ, si su función de probabilidad es de la forma

$$P(Y = y) = \frac{\lambda^y \, e^{-\lambda}}{y!}; \quad y = 0, 1, 2, \cdots$$

con esperanza y varianza

$$E(Y) = \mu = \lambda \qquad \text{y} \qquad Var(Y) = \sigma^2 = \lambda$$

Usando la notción corta, $Y \sim Poisson(\lambda)$.

Dado una sucesión de n variables aleatorias independientes $Y_1, Y_2, \cdots, Y_n$ con distribución Poisson, y parámetro λ_i, para $i = 1, 2, \cdots, n$, esto es, si

$$Y_i \sim Poisson(\lambda_i), \quad i = 1, 2, \cdots, n$$

entonces, se tienen los siguientes resultados:

a) La suma de las n variables aleatorias independientes Poisson con parámetro λ_i, es una Poisson con parámetro $\lambda = \sum \lambda_i$. Esto es,

$$\sum_{i=1}^{n} Y_i \sim Poisson(\lambda); \quad \text{donde} \quad \lambda = \sum_{i=1}^{n} \lambda_i \tag{1.11}$$

b) La distribución condicional de cada variable aleatoria poisson Y_i, dado que la suma de las variables es m y fija, es una binomia con parámetros m y π. Esto es,

$$Y_k \mid \sum_{i=1}^{n} Y_i = m \sim B\,(m,\ \pi_k)\,; \quad k = 1, 2, \cdots, n \tag{1.12}$$

con

$$\pi_k = \frac{\lambda_k}{\lambda} = \frac{\lambda_k}{\sum_{i=1}^{n} \lambda_i}; \quad k = 1, 2, \cdots, n$$

c) La distribución conjunta de las n variables aleatorias independientes Poison , dado la suma es n y fija, es una multinomial. Esto es,

$$(Y_1, Y_2, \cdots, Y_n) | \sum_{i=1}^{n} Y_i = m \ \sim \ Mult \left(n, \ \frac{\lambda_1}{\sum_{i=1}^{n} \lambda_i}, \cdots, \frac{\lambda_n}{\sum_{i=1}^{n} \lambda_i} \right) \tag{1.13}$$

La demostración de estos tres resultados no son difíciles, dado que corresponden a ejercicios de un curso intermedio de cálculo de probabilidades, pero, son de gran importancia para el análisis de tablas de contingencia bajo los modelos log-lineal Poisson.

1.5. Tablas de contingencia y el modelo Poisson

Las distribuciones binomial y multinomial son las distribuciones más usadas para el estudio de una tabla de contingencia, cuando el número de unidades clasificadas en la tabla de contingencia es limitado y fijo, con probabilidades de clasificación en las categorías no son tan pequeñas. Por ejemplo, en una encuesta política, podemos clasificar a los electores por sexo o grupos de edad y su preferencia hacia los candidatos en la contienda y estudiar la relación (independencia o asociación) entre los factores de clasificación.

En la práctica, se pueden presentar situaciones donde las unidades a ser clasificadas es grande ($n \to \infty$ ó n no es fijo, sino aleatorio) y las probabilidades de clasificación son pequeñas o corresponden a la ocurrencia de eventos raros. En este caso, la distribución Poisson es un modelo apropiado.

Sea una tabla de contingencia de dimensión $R \times S$, obtenido al clasificar las n unidades de la muestra de acuerdo a las categorías del factor fila o variable categórica X_A y el factor columna o variable categórica X_B, obteniéndose las contadas $y_{11}, y_{12}, \cdots, y_{RS}$, asociadas con las variables aleatorias $Y_{11}, Y_{12}, \cdots, Y_{RS}$, respectivamente, que son n variables aleatorias independientes Poisson con parámetro λ_{ij}. Esto es,

$$Y_{ij} \sim Poisson(\lambda_{ij}); \quad \text{para} \quad i = 1, 2, \cdots, R; \; j = 1, 2, \cdots, S$$

de modo que las probabilidades de clasificación en las celdas (i, j), son

$$\begin{aligned} \pi_{ij} &= P(Y_{ij} = y_{ij}) = P(Y = n_{ij}) \\ &= \frac{\lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij})}{y_{ij}!}; \quad \text{para} \;\; i = 1, 2, \cdots, R \;\; \text{y} \;\; j = 1, 2, \cdots, S \end{aligned}$$

Considerando que las contadas en cada una de las celdas de la tabla de contingencia ocurren como efecto de la realización de variables independientes, y como una suma de dichos efectos, la distribución Poisson son adecuados para analizar dichas ocurrencias que dan origen a las contadas en cada una de las celdas. Asimismo, de acuerdo a Agresti (2002)[5], los modelos log-lineal para tablas de contingencia están relacionados con el análisis ANOVA para variables categóricas, la distribución Poisson es un modelo apropiado para describir las probabilidades de las contadas $Y_{ij} = n_{ij}$, debido a que:

- Las contadas de cada celda son realización independiente de la variable aleatoria Y_{ij} Poisson, con μ_{ij} ocurrencias por intervalo de tiempo fijo.
- Si cada realización es un ensayo binomial con probabilidad de éxito $\pi_{ij} \to 0$, como son los eventos raros, conforme el número de ensayos se hace grande $n \to \infty$ y $n\pi_{ij} = \mu_{ij}$, la distribución límite es una Poisson con parámetro μ_{ij}. Esto es, $Y_{ij} \to Poisson(\mu_{ij})$, con $E(Y_{ij}) = \mu_{ij}$.

- Si las contadas son el resultado de un proceso estocástico Poisson, con media $\mu_{ij} = \lambda_{ij}t$, donde λ_{ij} es la razón de ocurrencia del evento en un intervalo de tiempo $(0, t]$ fijo.

Bajo este criterio, las contadas de cada celda ocurren de manera aleatoria y por tanto el total, también, es aleatorio. Esto es, el tamaño de la muestra no es fija, sino, aleatoria.

Bajo estas consideraciones, la distribución conjunta de las $R \times S$ celdas de una tabla de contingencia es

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}) &= \prod_{i=1}^{R} \prod_{j=1}^{S} P(Y_{ij} = y_{ij}) \\ &= \prod_{i=1}^{R} \prod_{j=1}^{S} \frac{\mu_{ij}^{y_{ij}} \cdot e^{-\mu_{ij}}}{y_{ij}!} \end{aligned}$$

Dado que las contadas $n_{ij} = y_{ij}$ son conocidas, el logaritmo de la función de verosimilitud de la muestra, es

$$\mathbf{L}((\mu) = \sum_{i=1}^{R} \sum_{j=1}^{S} \{y_{ij} \cdot \log(\mu_{ij}) - \mu_{ij} - y_{ij}!\} \tag{1.14}$$

2. El modelo lineal generalizado Poisson

Los modelos lineales generalizados fueron propuestos por Nelder y Wederburn[8] en 1972, quienes muestran que toda distribución que pertenece a la familia exponencial a un parámetro, puede ser modelado como un modelo lineal generalizado con la componente del error distinto a la normal, como son las distribuciones binomial, multinomial, Poisson, binomial negativa y otros. A partir de este artículo, los

modelos lineales generalizados se hacen populares para el estudio de la relación de variables no normales, consolidándose con el libro de MacCullag y Nelder (1989)[7].

2.1. Componentes del modelo lineal generalizado Poisson

Todo modelo lineal generalizado tiene tres principios, denominados componentes, que son la componente aleatoria, la componente sistemática y la función de enlace o link.

La componente aleatoria:

Sea Y una variable aleatoria con función de densidad (de distribución) perteneciente a una familia exponencial a un parámetro, de la forma

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

donde θ es el parámetro canónico y ϕ el parámetro de dispersión, para $\phi > 0$, y las funciones $a(\cdot)$, $b(\cdot)$ y $c(\cdot,\,\cdot)$ son monótonas y conocidas, tal que

$$E(Y) = \mu = \frac{\partial b(\theta)}{\partial \theta} \quad \text{y} \quad Var(Y) = \phi\frac{\partial^2 b(\theta)}{\partial \theta^2} = \phi\frac{\partial \mu}{\partial \theta} = \phi V(\mu)$$

donde $V(\mu)$ es llamado función de varianza.

Si los resultados anteriores lo aplicamos al caso de la distribución poisson, esto es, si $Y \sim Poisson(\mu)$, con

$$\begin{aligned} P(Y=y) &= \frac{\mu^y \exp\{-\mu\}}{y!} \\ &= \exp\left\{y \cdot \log(\mu) - \mu - log(y!)\right\} \end{aligned}$$

con parámetro canónico $\theta = \log(\mu)$, la función $b(\theta) = \exp\{\theta\}$, el parámetro de

dispersión $\phi = 1$ y $a(\phi) = 1$, y finalmente, $c(y, \phi) = -\log(y!)$. Además, la esperanza y la varianza de Y, es

$$E\,(Y) = \mu = \frac{\partial \exp\{\theta\}}{\partial \theta} = \exp\{\theta\} \quad \text{y} \quad Var(Y) = \phi \frac{\partial \mu}{\partial \theta} = \exp\{\theta\}$$

La componente sistemática:

Está restringido al predictor lineal de las variables explicativas $\mathbf{X} = (X_1, X_2, \cdots, X_p)$, de la forma

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

de la variable respuesta Y.

La función de enlace o link:

Está dado por la función $g(\mu)$ llamado función de enlace canónico de la distribución de probabilidades de Y, tal que conecta la esperanza μ de la distribución de la variable respuesta Y con el predictor lineal η. Esto es, si $E(Y|x_1, x_2, \cdots, x_n) = \mu_{Y|\mathbf{x}}$, la función de enlace es

$$g(\mu_{Y|\mathbf{x}}) = \eta$$

Podemos observar que, en un modelo lineal clásico, la componente aleatoria y la componente sistemática van juntos, con link identidad.

En el caso de la distribución poisson, el link canónico es el parámetro canónico $\theta = \log(\mu)$, tal que

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{2.2}$$

Esta ecuación, en la bibliografía estadística, es conocida como el predictor lineal del modelo lineal generalizado Poisson, y en muchos casos se le llama regresión de Poisson.

2.2. Estimación en la regresion de Poisson

Sean $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$ una muestra aleatoria *iid*, con distribución $Poisson(\mu_i)$, definido en (2.1). La distribución conjunta de la muestra está dado por

$$f(\mathbf{y}|\boldsymbol{\mu}) = \prod_{i=1}^{n} f(y_i|\mu_i) = \prod_{i=1}^{n} \exp\{y_i \cdot \log(\mu_i) - \mu_i - log(y_i!)\}$$

Si los valores de la muestra son observadas, tenemos las contadas observadas $\mathbf{Y}(\omega) = (y_1, y_2, \cdots, y_n)$, y se tiene la función de verosimilitud de la muestra, donde

$$\mathcal{L}(\boldsymbol{\mu}) = \prod_{i=1}^{n} \exp\{y_i \cdot \log(\mu_i) - \mu_i - log(y_i!)\} \tag{2.3}$$

El estimador $\hat{\boldsymbol{\mu}}$ es el estimador de máxima verosimilitud para $\boldsymbol{\mu}$, si $\hat{\boldsymbol{\mu}}$ maximiza la función de verosimilitud (2.3). Dado que la función de verosimilitud $\mathcal{L}(\boldsymbol{\mu})$ y su logaritmo $\log(\mathcal{L}(\boldsymbol{\mu})) = L(\theta, \phi)$ tienen un máximo en el mismo punto, para obtener el estimador de máxima verosimilitud para $\boldsymbol{\mu}$, será suficiente maximizar el logaritmo de la función de verosimilitud de la muestra, que es

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{n} \{y_i \cdot \log(\mu_i) - \mu_i - log(y_i!)\} \tag{2.4}$$

Considerando que cada observación y_i de la variable respuesta está asociado con las variables explicativas $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ a través de la función de enlace, para obtener el estimador $\hat{\boldsymbol{\mu}}$ para $\boldsymbol{\mu}$, se requiere introducir en el modelo el predictor lineal (2.2),

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

y

$$\mu_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

de modo que las estimaciones de los parámetros del predictor lineal

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$$

de acuerdo a McCullagh y Nelder[7] y Paula G.[13], se tiene que si $p \leq n$, la solución de la primera derivada del logaritmo de la función de verosimilitud (2.4) existe. Considerando que la estimación de $\boldsymbol{\mu}$ implica la estimación de $\boldsymbol{\beta}$, se tiene la siguiente notación del logaritmo de la función de verosimilitud:

$$L(\boldsymbol{\mu}) = L(\boldsymbol{\mu}|x_1, x_2, \cdots, x_n) = L(\boldsymbol{\beta})$$

Luego, usando la regla de la cadena, la función de score de la muestra, es

$$\begin{aligned} U(\beta_j) = \frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}) &= \frac{\partial L(\boldsymbol{\beta})}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \left\{ y_i \cdot \frac{1}{\mu_i} - 1 \right\} \cdot \frac{\partial \mu_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \{ y_i x_{ij} - \mu_i x_{ij} \} = 0 \end{aligned} \tag{2.5}$$

dado que

$$\begin{aligned} \frac{\partial \mu_i}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j} \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\} \\ &= \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\} \cdot x_{ij} \end{aligned} \tag{2.6}$$

Luego, como lo explica McCullagh y Nelder[7], las ecuaciones de máxima verosimilitud para β_j están dados para cada x_{ij}, con $j = 1, 2, \cdots, p$. De la estructura de La solución del sistema de ecuaciones no es cerrada, por lo que los aproximaremos usando métodos numéricos como el de Newton-Rapson, cuya forma general para $f(x) = 0$, la solución aproximada se obtiene mediante el algoritmo

$$x^{n+1} = x^n - \frac{f(x^n)}{f'(x^n)}$$

donde x es el vector de parámetros del predictor lineal η, $f(x)$ es la función score y $f'(x)$ es la derivada del la función score, cuya esperanza es la matriz de información de Fisher. Esto es,

$$i(\boldsymbol{\beta}) = -E\left[\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] = -E\left[\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right]$$

de modo que el algoritmo de Newton-Rapson para aproximar el valor del estimador para $\boldsymbol{\beta}$, es:

$$\hat{\boldsymbol{\beta}}^{(n+1)} = \hat{\boldsymbol{\beta}}^{(n)} - \left[\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1} \cdot U(\boldsymbol{\beta})\Bigg|_{\hat{\boldsymbol{\beta}}^{(n)}} \tag{2.7}$$

siendo $\hat{\boldsymbol{\beta}}^{(0)}$ el valor inicial y $\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ una matriz hessiana, con términos en (j,k):

$$\frac{\partial}{\partial \beta_k} U(\boldsymbol{\beta}, \phi) = \frac{\partial^2}{\partial \beta_k \beta_j} L(\theta, \phi | \boldsymbol{\beta}) = -\sum_{i=1}^{n} \mu_i \, x_{ij} x_{ik}$$

3. Modelos Log-lineal Poisson

Los modelos log-lineal son una clase de los modelos lineales generalizados, donde se desea estudiar las asociaciones entre dos o más variables categóricas, sin distinguir quién es la independiente y la dependiente o respuesta. A través de las contadas observadas n_{ij} y las contadas esperadas m_{ij} se desea estudiar la asociación que existe entre las variables individuales y sus interacciones de manera equivalente al análisis de ANOVA para las variables categóricas, donde el tamaño de una contada n_{ij} de una celda depende de las categorías o niveles de las variables de clasificación.

3.1. Modelos log-lineal Poisson para tablas de dos vías

El análisis log-lineal para tablas de contingencias de dos vías, es sencillo, destacando las tablas 2×2, donde el problema de la independencia se enfoca principalmente por la estructura de los odds ratios o razón de chances. En nuestro trabajo, el enfoque es general, para tablas $R \times S$, donde se presentan solamente dos modelos que son los modelos de independencia y los modelos con interacción, y su análisis requiere un número considerable de observaciones por celda, con $n_{ij} > 5$ y evitando celdas con cero contadas.

3.1.1. Tablas de contingencia de dos vías

Una tabla de contingencia de dos vías permite clasificar una serie de unidades u observaciones de acuerdo a dos variables categóricas, con el interés de estudiar la asociación entre dichas variables. Esto es, dado una tabla de contingencia asociada con las variables categóricas X_A y X_B, de dimensión $R \times S$, con probabilidades de clasificación π_{ij} en la celda (i, j), tal que

$$\pi_{ij} \geq 0 \qquad \text{y} \qquad \sum_{i=1}^{R} \sum_{j=1}^{S} \pi_{ij} = 1$$

con probabilidades marginales y total

$$\sum_{j=1}^{S} \pi_{ij} = \pi_{i+} \qquad \sum_{i=1}^{R} \pi_{ij} = \pi_{+j} \qquad \sum_{i=1}^{R} \sum_{j=1}^{S} \pi_{ij} = \pi_{++}$$

y contadas esperadas $m_{ij} = \mu_{ij}$, con las mismas características descritas para las probabilidades marginales y el total.

Consideremos que una muestra de n unidades de la población fueron clasificadas en una tabla de contingencia de dos vías, de dimensión $R \times S$, obteniéndose las contadas y_{ij}, tal que:

- Las contadas asociadas con cada una de las $R \times S$ celdas, tienen distribución

multinomial, con probabilidades de clasificación π_{ij}, para $i = 1, 2, \cdots, R$; y $j = 1, 2, \cdots, S$.

- Cada una de las contadas son realizaciones de una variable aleatoria Poisson con media $\lambda_{ij} = \mu_{ij}$. Esto es,

$$Y_{ij} \sim Poisson(\mu_{ij}) \text{ para } i = 1, 2, \cdots, R \text{ y } j = 1, 2, \cdots, S$$

- las $R \times S$ contadas de la tabla de contingencia son realizaciones independientes de las variables aleatorias Poisson con media μ_{ij}, tal que

$$Y_{++} = \sum_{i=1}^{R}\sum_{j=1}^{S} Y_{ij} \ \sim \ Poisson\left(\mu_{++} = \sum_{i=1}^{R}\sum_{j=1}^{S} \mu_{ij}\right)$$

- Dado que $\sum_{i=1}^{R}\sum_{j=1}^{S} y_{ijk} = n$, la distribución condicional de las $R \times S$ contadas es una multinomial con probabilidades de clasificación $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$. Esto es,

$$Y_{ij} | \sum_{i=1}^{R}\sum_{j=1}^{S} y_{ijk} = n \ \sim \ M\left(n, \ \pi_{ij} = \frac{\mu_{ij}}{\mu_{++}}\right)$$

  para $i = 1, 2, \cdots, R$ y $j = 1, 2, \cdots, S$.

- Si $\hat{\mu}_{ij}$ es el estimador de máxima verosimilitud para las contadas esperadas μ_{ij}, entonces

$$\sum_{i=1}^{R}\sum_{j=1}^{S} y_{ij} = \sum_{i=1}^{R}\sum_{j=1}^{S} \hat{\mu}_{ij} = n$$

3.1.2. Modelos log lineal para tablas de dos vías

Bajo las consideraciones anteriores y siguiendo a Christensen[10], Fienberg[12] y Agresti[5] y otros autores, para el análisis de la asociación de las variables de una tabla de contingencia de dos vías, existen dos modelos:

(1) **Modelo de independencia completa:**

La hipótesis de independencia de los dos factores, es

$$H_0 : \ \pi_{ijk} = \pi_{i+} \cdot \pi_{+j}$$

Bajo la hipótesis nula, el estimador de máxima verosimilitud para m_{ij} es

$$\hat{m} = \frac{n_{i+} \cdot n_{+j}}{n}$$

con el que podemos calcular la estadística de test de la chi-cuadrado de Pearson y el test de razón de verosimilitud, que tienen distribución chi-cuadrada con $(R-1)(S-1)$ grados de libertad.

Bajo el modelo de independencia de los factores, el modelo log-lineal de independencia, es

$$M^1 : \quad \log(m_{ij}) = u + u_i^A + u_j^B$$

tal que, como son desviaciones respecto a la media u,

$$\sum_i^R u_i^A = \sum_i^R u_j^B = 0$$

(2) **El modelo saturado:**

Bajo el modelo de muestreo multinomial se tiene que $m_{ij} = n_{++}\pi_{ij}$ y en el producto multinomial, se tiene que $m_{ij} = n_{i+}\pi_{ij}$ y el modelo log-lineal bajo el modelo mde independencia, podemos adicionar un término de interacciónl, obteniendo

$$M^2 : \quad \log(m_{ij}) = u + u_i^A + u_j^B + u_{ij}^{AB}$$

tal que

$$\sum_i^R u_i^A = \sum_i^R u_j^B = 0 \quad \text{y} \quad \sum_i^R u_{ij}^{AB} = \sum_i^R u_{ij}^{AB} = 0$$

3.2. Modelos log-lineal Poisson para tablas de tres vías

3.2.1. Tablas de contingencia de tres vías

El análisis de tablas de contingencia de tres vías, consiste en analizar la asociación de tres variables categóricas: fila, columna y profundidad, considerando las hipótesis de independencia total, parcial y condicionada, entre las variables.

Formalizando, consideremos las tres variables categóricas: X_A la variable fila, X_B la variable columna y X_C la variable profundidad, con categorías R, S y T, respectivamente, de modo que la dimensión de la tabla es $R \times S \times T$, con probabilidades de clasificación π_{ijk} en la celda (i, j, k), con características equivalentes de una tabla de contingencia de dos vías.

Consideremos que una muestra de n unidades de la población fueron clasificadas en una tabla de contingencia de tres vías, de dimensión $R \times S \times T$, obteniéndose las contadas y_{ijk}, tal que:

- Las contadas asociadas con cada una de las $R \times S \times T$ celdas, tienen distribución multinomial, con probabilidades de clasificación π_{ijk}, para $i = 1, 2, \cdots, R$; $j = 1, 2, \cdots, S$ y $k = 1, 2, \cdots, R$.
- Cada una de las contadas son realizaciones de una variable aleatoria Poisson con media $\lambda_{ijk} = \mu_{ijk}$. Esto es,

$$Y_{ijk} \sim Poisson(\mu_{ijk})$$

  para $i = 1, 2, \cdots, R$; $j = 1, 2, \cdots, S$ y $k = 1, 2, \cdots, R$.
- las $R \times S \times T$ contadas de la tabla de contingencia son realizaciones independientes de las variables aleatorias Poisson con media μ_{ijk}, tal que

$$Y_{+++} \sim Poisson\,(\mu_{+++})$$

con

$$Y_{+++} = \sum_{i=1}^{R}\sum_{j=1}^{S}\sum_{k=1}^{T} Y_{ijk} \quad \text{y} \quad \mu_{+++} = \sum_{i=1}^{R}\sum_{j=1}^{S}\sum_{k=1}^{T} \mu_{ijk}$$

- Dado que $\sum_{i=1}^{R}\sum_{j=1}^{S}\sum_{k=1}^{T} y_{ijk} = n$, la distribución condicional de las $R \times S \times T$ contadas es una multinomial con probabilidades $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$.
- Si $\hat{\mu}_{ijk}$ es el estimador de máxima verosimilitud para μ_{ijk}, entonces

$$\sum_{i=1}^{R}\sum_{j=1}^{S}\sum_{k=1}^{T} y_{ijk} = \sum_{i=1}^{R}\sum_{j=1}^{S}\sum_{k=1}^{T} \mu_{ijk} = n$$

3.2.2. Modelos log lineal para tablas de tres vías

Bajo las consideraciones anteriores y siguiendo a Christensen[10], Fienberg[12] y Agresti[5] y otros autores, para el análisis de una tabla de contingencia de tres vías existen ocho modelos:

(1) **Los tres factores independientes o modelo de independencia completa:**

$$H_0 : \ \pi_{ijk} = \pi_{i++} \cdot \pi_{+j+} \cdot \pi_{++k}$$

obteniéndose el modelo

$$M^1 : \quad \log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C$$

donde el test de la chi-cuadrado y el test de razón de verosimilitud son evaluados con $RST - R - S - T + 2$ grados de libertad.

(2) **Modelos donde un factor es independiente de los otros dos:**

El factor fila es independiente de los factores columna y profundidad

$$H_0 : \ \pi_{ijk} = \pi_{i++} \cdot \pi_{+jk}$$

El modelo loglineal

$$M^2 : \quad \log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{jk}^{BC}$$

El factor columna es independiente de los factores fila y profundidad

$$H_0 : \ \pi_{ijk} = \pi_{+j+} \cdot \pi_{i+k}$$

El modelo loglineal

$$M^3 : \quad \log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ik}^{AC}$$

El factor profundidad es independiente de los factores fila y columna

$$H_0 : \ \pi_{ijk} = \pi_{++k} \cdot \pi_{ij+}$$

El modelo loglineal

$$M^4 : \quad \log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB}$$

(3) **Modelos donde independencia condicional:**

Dado el factor profundidad, el factor fila y columna son independientes

$$H_0 : \ \pi_{ijk} = \pi_{i+k} \cdot \pi_{+jk} | \pi_{++k}$$

El modelo loglineal

$$M^5 : \quad \log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ik}^{AC} + u_{jk}^{BC}$$

Dado el factor columna, el factor fila y profundidad son independientes

$$H_0 : \ \pi_{ijk} = \pi_{ij+} \cdot \pi_{+jk} | \pi_{+j+}$$

El modelo loglineal

$$M^6 : \quad \log\left(m_{ijk}\right) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{jk}^{BC}$$

Dado el factor fila, el factor columna y profundidad son independientes

$$H_0 : \ \pi_{ijk} = \pi_{ij+} \cdot \pi_{i+k} | \pi_{i++}$$

El modelo loglineal

$$M^7 : \quad \log\left(m_{ijk}\right) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + + u_{ik}^{AC}$$

(4) **El modelo saturado:**

Christensen[10] afirma que este modelo fue estudiado por Bartlett (1935) y puede ser expresado en términos de los odds ratios, en el sentido que los odds ratios son la misma para cualquier índice de la profundidad, esto es, en el caso particular

$$M^8 : \quad \frac{\pi_{111}\pi_{ij1}}{\pi_{i11}\pi_{1j1}} = \frac{\pi_{11k}\pi_{ijk}}{\pi_{i1k}\pi_{1jk}}$$

para $i = 1, 2, \cdots, R$; $j = 1, 2, \cdots, S$; $k = 1, 2, \cdots, T$.

El modelo loglineal

$$M^8 : \quad \log\left(m_{ijk}\right) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + + u_{ik}^{AC} + u_{jk}^{BC} + + u_{ijk}^{ABC}$$

Los grados de libertad de los modelos log lineal de tres vías, se obtienen por la combinación de los grados de libertad de los términos del modelo, que de acuerdo a Christensen[10] y Fienberg[12], son

Término	Grados de libertad
u	1
u^A	$R-1$
u^B	$S-1$
u^C	$T-1$
u^{AB}	$(R-1)(S-1)$
u^{AC}	$(R-1)(T-1)$
u^{BC}	$(S-1)(T-1)$
u^{ABC}	$(R-1)(S-1)(T-1)$

El ajuste de los datos a los modelos lo haremos utilizando el software estadístico R, partidendo del modelo saturado y llegando al modelo de independencia o viceversa, buscando el mejor ajuste.

Un problema que se presenta en el análisis de tablas de contingencia es que si la muestra es muy pequeña, donde las celdas contienen cero contadas o contadas de cinco o menos los problemas de estimación bajo el modelo log-lineal se hace dificil o distorcionan los resultados. Otro caso extremo es la estimación bajo el modelo saturado, donde se requieren de los métodos numéricos para aproximar el ajuste, debido a que el problema de estimación no tiene una solución cerrada o simple. En estos casos, es de mucha ayuda usar los métodos de remuestreo, como el bootstrap.

4. El método de remuestreo bootstrap

La técnica del bootstrap, es una técnica estadística perteneciente a la clase de los procedimientos de remuestreo a partir de un conjunto de datos originales. En esta clase de modelos de remuestreo se tiene el Jacknife propuesto por Quenouille (1949), que dado un estimador $\hat{\theta}_n$ obtenido usando los n datos de la muestra, el estimador Jacknife es el mismo estimador evaluado con solo $n-k$ datos, que los denotaremos como $\hat{\theta}_{n-k}$. Este estimador lo introdujo Tukey, en 1958, en la técnica del análisis

exploratorio de datos - EDA, para $k = 1$, como una medida de la influencia de cada una de las observaciones excluidas i en la formación del valor del estimador $\hat{\theta}_n$, seguidamente, la técnica del Jaclnife se amplía como una técnica multipropósito para evaluar la estabilidad de la varianza y el sesgo de un estimador en un proceso de prueba de hipótesis.

Con la ayuda del computador para los cálulos, Efron (1979)[14] propone el bootstrap como un método alternativo de remuestreo al Jacknife para aproximar el valor posible de los parámetros de la distribución de un estimador $\hat{\theta}_n$. En los siguientes trabajos de Efron, con lo siguientes trabajos que publica Efron, como son Efron (1981)[18], Efron (1983)[15], Efron y Tibshirani (1993)[16] y Efron (1993)[17] unifican las ideas del remuestreo y proponen la técnica del Bootstrap como una metodología estadística para el cálculo del error estándar y el sesgo de un estimador usando muestras independientes, en situaciones donde el proceso de estimación del parámetro es es compleja y su aproximación por métodos estándares no es apropiada. Pero, los métodos bootstrap, si bien eran interesantes desde el punto de vista técnico, requerían de un gran esfuerzo computacional o uso intensivo de la computadora, con el propósito de relajar algunas de las condiciones de la tradicional inferencia estadística con propósitos de hacer inferencias.

En la actualidad, la aplicación de la técnica del Bootstrap es amplio, en casi todas las áreas de la estadística: los modelos de regresión, los modelos lineales generalizados, los problemas de clasificación, etc. y el muchas disciplinas del conocimiento: la , la biología y medicina, la psicología, etc.

4.1. El método Bootstrap

Sea $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ una muestra aleatoria de tamaño n de una población Y, que constituyen n variables aleatorias *iid* con función de distribución F_θ, esto es,

$$Y \sim F_\theta; \quad \text{con} \quad \theta \in \Theta \tag{4.1}$$

A partir de la información contenida en la muestra, de acuerdo con (Efron 1979), podemos enfrentar con los problemas de la inferencia estadística:

(1) La determinación del valor de un estimador de un parámetro de interés y la evaluación de la precisión del estimador mediante el error estándar.

(2) La determinación de intervalos de confianza para el parámetro de interés.

(3) Realizar contrastes de hipótesis a cerca del parámetro de interés.

Estos problemas los podemos solucionar bajo el paradigma:

- De la estadística paramétrica, donde la distribución F_θ es conocida y el parámetro θ no es conocido.

  Sea $\hat{\theta}$ el estimador del parámetro θ asociado con la distribución F, calculado a partir de la muestra original, que sustituyendo el estimador en F obtendremos la distribución $\hat{F}$, que puede ser utilizada para generar sub-muestras aleatorias, con las que podemos hacer las estimaciones Bootstrap de interés.

- De la estadística no paramétrica, donde se asume que la distribución F_θ es no conocida.

  Dado que no se conoce F_θ, lo podemos estimar por la distribución empírica $\hat{F}_n$, con probabilidad de masa $1/n$ para cada punto muestral. En este caso, las remuestras las obtendremos de la muestra original, generando sub-muestras con reemplazo de tamaño n, con las que obtendremos los estimadores Bootstrap.

Efron, en sus diferentes trabajos, introduce el Bootstrap como una técnica estadística para solucionar problemas de inferencia estadística cuando no se conoce el estimador del error estándar del estimador o cuando su estimación es compleja y su aproximación por métodos numéricos conocidos no es apropiada.

4.2. Algoritmo del método Bootstrap

El algoritmo de aplicación de la técnica del Bootstrap es la siguiente:

1) Sea $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ una muestra de tamaño n de una población Y con función de distribución F_θ, esto es,

$$Y \sim F_\theta; \quad \text{con} \quad \theta \in \Theta \tag{4.2}$$

que la llamaremos muestra original, con el que calculamos el valor del estimador

$$\hat{\theta} = T(y_1, y_2, \cdots, y_n) = T(F_\theta) \tag{4.3}$$

Una limitación, para que el Bootstrap nos proporcione resultados razonables, es que la muestra original debe ser lo suficientemente grande, de modo que sea representativa de la población que dió origen a la muestra original.

2) Se generan las M sub-muestras de tamaño n de la muestra original mediante el muestreo con reemplazo y se calculan las estadísticas de interés. Esto es,

$$\mathbf{y}_i^* = (y_{i1}^*, y_{i2}^*, \cdots, y_{in}^*) \Longrightarrow \hat{\theta}_i = T(\mathbf{y}_i^*); \quad i = 1, 2, \cdots, M \tag{4.4}$$

3) Cálculo de los estimadores Bootstrap:

- Cálculo del estimador del parámetro θ

$$\hat{\theta}_{BOOT} = \frac{\sum_{i=1}^{M} \hat{\theta}_i}{M} \tag{4.5}$$

- Cálculo del error estándar del estimador $\hat{\theta}$

$$EE(\hat{\theta})_{BOOT} = \left\{ \frac{1}{M(M-1)} \sum_{i=1}^{B} \left(\hat{\theta}_i - \hat{\theta}_{BOOT} \right)^2 \right\}^{1/2} \tag{4.6}$$

4.3. Intervalos de confianza Bootstrap

En un proceso de inferencia estadística, la estimación puntual $\hat{\theta}$ del parámetro θ no presenta la medida de los componentes del error cuadrático medio: la precisión del estimador, que está asociado con la variabilidad o dispersión del estimador, y la exactitud de la estimación, que está asociado con el sesgo.

Los intervalos de confianza incorporan las deficiencias de la estimación puntual, donde los intervalos de confianza Bootstrap los podemos obtener utilizando diferentes criterios o estrategias, que para nuestros requerimientos, como sugiere Efron y Tibshirani (1993)[16], utilizaremos el método percentil y el método pivotal, cuya forma básica de un intervalo de confianza bootstrap, es de la forma

$$\hat{\theta} - \epsilon \leq \theta \leq \hat{\theta} + \epsilon$$

donde

1. Los intervalos de confianza Bootstrap por el método de percentiles, se obtiene estimando la función distribución F del estimador $\hat{\theta}$, de modo que el intervalo de confianza de nivel $(1-\alpha)$ para el parámetro de interés está dado por

   $$\left[\hat{F}^{-1}(\alpha/2),\ \hat{F}^{-1}(1-\alpha/2)\right]$$

   donde, $\hat{F}^{-1}(\alpha)$ es el percentil α de la distribución del estimador Bootstrap $\hat{\theta}$.

2. Los intervalos de confianza Bootstrap por el método pivotal o t-Bootstral, son construidos a partir de la aproximacion de la transformacion $N(0,1)$, donde

   $$Z = \frac{\hat{\theta} - \theta}{EE(\hat{\theta})} \sim t_{\alpha/2,(n-1)gl}$$

   de modo que el intervalo de confianza es de la forma

   $$\hat{\theta} \pm t_{\alpha/2,(n-1)gl}\ \hat{EE}(\hat{\theta})$$

de modo que el valor de z es estimado directamente de los datos muestrales en el re-muestreo,

$$z_i^* = \frac{\hat{\theta}_i^* - \hat{\theta}}{\hat{EE}(\hat{\theta}_i^*)}; \quad i = 1, 2, \cdots, M \tag{4.7}$$

4.4. El bootstrap y los modelos lineales generalizados

Crawley (2007)[19] afirma que el uso de los modelos lineales generalizados es adecuado, cuando el modelo lineal no satisface el supuesto de la homocedasticidad o la varianza de la variable respuesta no es constante, y no satisfacen el supuesto de normalidad de los errores o si los errores no son normales. Estos dos problemas se presentan cuando se analizan datos de tablas de contingencia, más si trabajamos con el modelo de errores Poisson, donde la varianza se incrementa conforme la esperanza se incrementa y viceversa. Por tanto, el problema de la sobredispersión siempre estará presente.

Bajo el modelo Poisson, las observaciones $Y = y$ son contadas o números enteros no negativos ($y \geq 0$), cuya varianza varía conforme varía la media. Además, en análisis de una tabla de contingencia el interés es comparar las discrepancias entre las contadas observadas $y_{ij} = n_{ij}$ con las contadas esperadas m_{ij}, que son calculadas bajo el modelo de la hipótesis nula, donde la medida de la discrepancia es llamado bondad de ajuste Read y Cressie (1988)[21] y Winkler (1996)[22].

Un problema crucial que se presenta al contrastar una hipótesis es conocer la distribución verdadera del estimador $\hat{\theta}_n$, debido a que dicha distribución no es conocida o es de cálculo difícil. Una forma de aproximarnos a dicha distribución es mediante la distribución asintótica, donde tanto $\theta = E(\hat{\theta}_n)$ y el error estándard $EE(\hat{\theta}_n)$ los podemos aproximar mediante estimadores bootstrap a partir de un número de muestras independientes bootstrap muy grande ($n \to \infty$). Bajo este criterio, las estadísticas de bondad de ajuste: el test chi cuadrada de Pearson y el test de razón de

máxima verosimilitud tienen distribución asintótica chi-cuadrado con $(R-1)(S-1)$ grados de libertad. Este criterio lo podemos describir mediante el siguiente gráfico, el mismo que tomamos de Efron y Tibshirani (1993)[16] y de Winkler (1996)[22]

$$\begin{array}{ccccccc} F & \longrightarrow & \mathbf{Y} & \longrightarrow & \hat{F} & \longrightarrow & \mathbf{Y}^* \\ & & \downarrow & & & & \downarrow \\ & & G(\mathbf{Y}) & \longleftarrow & \mathcal{L}(G(F)) \approx \mathcal{L}(G(\hat{F})) & \longleftarrow & G(\mathbf{Y}^*) \end{array}$$

donde

- El vector aleatorio $\mathbf{Y} \sim F(\beta)$ y $F(\cdot)$ es la distribución del vector, que como en nuestro caso, es la distribución Poisson, y $\boldsymbol{\beta}$ es el vector de parámetros no conocidos de la distribución.
- La estadística de bondad de ajuste los designamos por $G(\mathbf{Y}) = G(F)$ cuya funcional es conocida, cuya distribución o ley de probabilidades es $\mathcal{L}(G(F))$ de interés.
- Si definimos el estimador de la distribución $F(\boldsymbol{\beta})$ por $\widehat{F(\boldsymbol{\beta})} = \hat{F}$, que puede ser calculado con los datos de la muestra, donde F es conocida (bootstrap paramétrico) o es la función de distribución empírica (bootstrap no-paramétrico), y $\hat{F}$ podría ser $F(\hat{\boldsymbol{\beta}})$.
- El estimador bootstrap consiste en aproximar la distribución o ley de probabilidades $\mathcal{L}(G(F))$ por $\mathcal{L}(G(\hat{F}))$.

Bajo las consideraciones anteriores, la tecnica del bootstrap en los modelos log-lineal Poisson los podemos aplicar desde los siguientes puntos de vista:

(a) Las pruebas de bondad de ajuste, relacionadas con las hipótesis de independencia o de asociación de las variables, usan las estadísticas de test de la chi-cuadrado de Pearson y la devianza o test de razón de verosimilitud, las

mismas que miden las discrepancias entre las contadas observadas y las contadas esperadas o teoricas. Pero, en la teoría estadística o estadística matemática se estudia que la estadística de la chi-cuadrado de Pearson es una estadística que, asintóticamente (cuando $n \to \infty$) tiene distribución chi-cuadrado.

Cuando n es pequeño, el supuesto asintótico no se cumple, más si las celdas de la tabla presentan ceros o son menores a 5 y el p-valor de las estadísticas son relativamente grandes (p-valor ≥ 0.10). En este caso, la estimación de las contadas esperadas y_{ij} y $\hat{m}_{ij}$ los podemos mejorar usando el Bootstrap no paramétrico, permitiendo disminuir el p-valor.

(b) El modelo log-lineal Poisson es definido por el predictor lineal (2.2):

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Si $\hat{\boldsymbol{\beta}}$ es el estimador de máxima verosimilitud para $\boldsymbol{\beta}$, entonces $\hat{\boldsymbol{\beta}}$ es asintóticamente normal. Esto es, si $n \to \infty$, se tiene que

$$\frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{EE(\hat{\boldsymbol{\beta}})} \ \to \ N(0,1); \quad n \to \infty$$

Si n es pequeño, la distribución asintótica normal del estimador de máxima verosimilitud $\hat{\boldsymbol{\beta}}$ puede ser pobre. En este caso, el error estándar del estimador puede ser mejorado usando el bootstrap paramétrico.

5. Materiales y métodos

La aplicación de las técnicas estadísticas desarrolladas en el presente trabajo de investigación lo haremos con los datos recolectados por un grupo de investigación del Servicio de Neumología del Hospital Nacional María Auxiliadora de San Juan de Miraflores sobre **secuelas de la tuberculosis pulmonar** en pacientes con antecedente de tuberculosis pulmonar por los Doctores investigadores De los Rios y Bravo (2012).

Debemos mencionar que los datos originales cedidos consistían de 58 observaciones, correspondientes a un mismo número de pacientes evaluados y tratados por tuberculosis y a quienes se les hizo un seguimiento para observar los signos y síntomas característicos de secuela. El objetivo del trabajo de De los Rios y Bravo (2012) fue hacer un análisis descriptivo de los datos.

En el presente trabajo de investigación, se consideró interesante analizar la relación entre las variables antecedentes y las variables síntomas de secuela, mediante tablas de contingencia de 2 vías y posteriormente, y posteriormente, se decidió considerar el análisis de tablas de contingencia de 3 vías, introduciendo en nuestro análisis las variables datos generales del paciente.

Un problema que se presentaron en el análisis de los datos, fue que el número de datos era muy pequeño, dado que las tablas de contingencia correspondientes, contenían celdas con cero contadas y contadas de 5 o menos, situación que complica el análisis y la calidad de los resultados. Ante esta situación, con el apoyo y asesoramiento de la Doctora Bravo se recolectó, a partir de las fichas clínicas de pacientes nuevos una serie de dados en dos Hospitales, de los cuales se validaron como buenos solo 27, con los cuales mejoró la estructura de las tablas de contingencia, considerando que el análisis se mejoraría, aun más, si el número de datos fuera superior a los 200.

5.1. Descripción del problema

De los Rios y Bravo (2012) afirman que *"la tuberculosis es una enfermedad infectocontagiosa que suele afectar predominantemente a los pulmones y es causada por una bacteria (Mycobacterium tuberculosis).*

En el Perú, la tasa de incidencia para el año 2010 fue de 96.1 por cada 100,000 habitantes lo que corresponde a 32,477 nuevos casos ese año.

Del 2001 al 2005 se han diagnosticado y tratado 177,988 casos de tuberculosis en todo el país. En el año 2005 la eficiencia alcanzada (% curación) es del 90 % de los cuales,

el 96 % fueron confirmados con frotis negativo al término del tratamiento. Por lo tanto, los esquemas primarios mantienen los niveles de eficiencia encontrándose valores por encima del 85 % señalado por la OMS.

La tuberculosis, debido a su carácter infeccioso y necrotizante produce efectos destructivos en el parénquima pulmonar y bronquial que persisten luego de la cura bacteriológica, dando lugar a las secuelas de la tuberculosis. Se postula que la magnitud de la secuela depende de la extensión del proceso previo.

Desde el punto de vista radiológico existen una serie de alteraciones estructurales en las porciones pulmonares y extra-pulmonares del tórax como consecuencia de la Tuberculosis que han sido clasificadas de la siguiente manera: Lesiones parenquimatosas (tuberculomas, cavidades de paredes delgadas, estériles, bandas cicatriciales, pulmones terminales, aspergiloma y asociación con carcinoma broncogénico), de la vía aérea (bronquiectasias, estenosis traquebronquial y broncolitiasis), vasculares (arteritis bronquial y pulmonar que incluye trombosis, dilatación de arterias bronquiales y aneurisma de Rasmussen), mediastinales (nódulos linfáticos calcificados, fístula esofágica, pericarditis constrictiva y mediastinitisfibrosante), pleurales (fibrotórax, fístula broncopleural y neumotórax) y de la pared torácica.

Debido a la gran variedad de secuelas anatómicas, podemos deducir que los cuadros clínicos que presentarán los pacientes post TBC serán de diversa índole y severidad. Esto condiciona diagnósticos errados por lo que son catalogados como EPOC, bronquiectasias, asma, hiper-reactividad bronquial o fibrosis pulmonar".

5.2. La muestra y operacionalización de las variables

Considerando la muestra de tamaño 85 datos observados en pacientes con diagnóstico de secuela, y en cada registro del paciente se observaron 36 variables (características del paciente diagnosticado), de los cuales, para los fines de nuestro trabajo de investigación, se tomaron 12 variables que los describimos a continuación:

(I) **Datos generales del paciente**

- **Sexo** Categorías: hombre (1), mujer (2).
- **Edad** Categorías: de 15 a 29 años (1), de 30 a 49 años (2), de 50 y más años (3).
- **Indice de masa corporal (imc)** Categorías: con imc menor a 18.50 (1), de 18.5 a 24.9 (2), de 25.0 y más años (3).

(II) **Antecedentes del paciente**

- **Tiempo de tratamiento antituberculoso (ttratam)** Categorías: de 0 a 6 meses (1), de 7 a 12 meses (2), de 13 a 18 meses (3), de 19 y más años (4).
- **Número de episodios (nepisod)** Categorías: número de veces que el paciente tuvo el diagnóstico de tuberculosis 1 episodio (1), 2 episodios (2), más de 2 episodios (3).
- **Diagnóstico de asma bronquial (dasma)** Categorías: si, previo a TBC (1), si, posterior a TBC (2), no (3).
- **Diagnóstico fibrosis pulmonar (dfibrosis)** Categorías: si, previo a TBC (1), si, posterior a TBC (2), no (3).
- **Diagnóstico bronquiectasias (bronqtsis)** Categorías: si, previo a TBC (1), si, posterior a TBC (2), no (3).

(III) **Síntomas de secuela en el paciente**

- **Dísnea MRCm (disnea)** de menor a mayor tolerancia a la actividad física.
  Categorías: nivel 1 (1), nivel 2 (2), nivel 3 (3), nivel 4 (4).
- **Tos crónica (toscr)** Categorías: expectoración mucoide (1), expectoración mucopurulenta (2), expectoración hemoptoica (3), tos seca (4).
- **Sibilancias (sibilanc)** Categorías: si (1), no (2).

- **Hemoptisis (hemoptosis)** Categorías: uno o más episodios/año (1), uno o más visitas emergencia/año (2), asociado a infecciones (3).

5.3. Análisis descriptivo univariado de las variables

Las característica univariadas de cada una de las variables, por sexo, consideradas en el trabajo de investigación, los presentamos en los siguientes cuadros:

(a) Datos generales del paciente:

	Hombre	Mujer	Total	%
edad				
15 a 29	3	5	8	9.4
30 a 49	5	19	24	28.2
50 y más	18	35	53	62.4
IMC				
menor a 18.5	19	39	58	68.2
18.5 a 24.9	1	14	15	17.6
25 y más	6	6	12	14.1
Total	26	59	85	100.0
%	30.6	69.4	100.0	

Los 2/3 de los pacientes diagnosticados con secuela son mujeres; los 2/3 son mayores de 50 años y el 90.6 % tienen una edad de 30 o más; el 68.2 % tienen un imc por debajo de 18.5. Por tanto, el grupo de nujeres mayores de 30 años son las más vulnerables por la TBC y las consecuencias de las secuelas.

(b) Antescedentes del paciente:

El 50 % de los pacientes con diagnóstico de TBC reciben un tratamiento de no más de 6 meses y 1/3 de entre 7 a 12 meses; 2/3 de los pacientes han tenido un episodio de TBC y 1/3 dos episodios; el 56.5 % de los pacientes con diagnóstoco de TBC

sufren del asma y 1/3 no tiene asma.

	Hombre	Mujer	Total	%
ttratam				
0 a 6	11	31	42	49.4
6 a 12	14	17	31	36.5
12 a 18	0	2	2	2.4
18 y más	1	9	10	11.8
nepisod				
1	18	35	53	62.4
2	8	19	27	31.8
3 y más	0	5	5	5.9
dasma				
previo tbc	15	33	48	56.5
posterior tbc	3	4	7	8.2
no	8	22	30	35.3
Total	26	59	85	100.0
%	30.6	69.4	100.0	
dfibrosis				
previo tbc	23	49	72	84.7
posterior tbc	0	0	0	0.0
no	3	10	13	15.3
bronqtsis				
previo tbc	22	42	64	75.3
posterior tbc	0	2	2	2.4
no	4	15	19	22.4
Total	26	59	85	100.0
%	30.6	69.4	100.0	

El 84.7 % de los pacientes tuvieron el diagnóstico de fibrosis pulmonar y el 75 % el diagnóstico de broquiestasis prebio a la TBC.

(c) Síntomas de secuela en el paciente:

	Hombre	Mujer	Total	%
disnea				
nivel-1	8	22	30	35.3
nivel-2	10	24	34	40.0
nivel-3	4	5	9	10.6
nivel-4	4	8	12	14.1
toscr				
espect-1	15	36	51	60.0
espect-2	3	14	17	20.0
espect-3	8	5	13	15.3
espect-4	0	4	4	4.7
sibilanc				
si	21	50	71	83.5
no	5	9	14	16.5
hemoptis				
nivel-1	8	27	35	41.2
nivel-2	15	22	37	43.5
nivel-3	3	10	13	15.3
Total	26	59	85	100.0
%	30.6	69.4	100.0	

Las consecuencias o secuelas que deja la TBC a un paciente que lo adquirió, son la sibilancia (83.5 %), la tos crónica en los dos niveles: espectoración mucoide (60 %) y espectoración mucopurulenta (20 %), la hemoptisis con uno o más episodios al año (41.2 %) y con una o más visitas a Emergencia al año (41.2 %) y la dísnea en los niveles 1 y 2 (35.3 % y 40.0 %), que limita a realizar esfuerzo al paciente.

5.4. Análisis log-lineal Poisson con tablas de dos vías

En esta sección analizaremos la relación entre las variables *antecedentes del paciente* y *síntomas de secuela del paciente*, considerando la hipótesis nula de independencia de los factores, esto es,

H_0 : los antecedentes son independientes de los síntomas

vs

H_1 : los síntomas están asociados a los antecedentes

Para contrastar la hipótesis nula utilizaremos la estadística de test de la chi-cuadrada de Pearson (chi.cuad) y el test de razón de verosimilitud (trv), cuyos resultados son equivalentes, los mismos que los presentamos en la siguiente tabla:

VA ⋱ VS	disnea	toscr	sibilanc	hemoptsis	nepisano
ttratm					
nepisod	XXX		XXX	XXX	XXX
dasma	XXX		XXX		XXX
dfibrosis				XXX	
broqtsis	XXX	XXX		XXX	

Donde, las celdas vacías nos muestran que el tes es no significativo al 5 %, en el sentido que existe independencia entre las variables. Las celdas de la tabla con tres aspas no muestra que existe asociación entre las variables, en el sentido que se rechaza la hipótesis nula. Esto es,

- La variable **tiempo de tratamiento** es independiente con todas las variables **síntomas**. Luego, la variable tiempo de tratamiento no influye en los síntomas de secuela, situación que parese ser muy razonable.
- la variable antecedente **número de episodios de TBC** está asociado a las variables síntomas **dísnea, sibilancia, hemoptosis y el número de epi-**

sodios año. Por tanto, este antecedente es importante en los síntomas de de secuela.

- La variable antecedente **diagnóstico de asma** está asociado con las variables síntomas **dísnea, sibilancia y el número de episodios año**.
- La variable antecedente **diagnóstico de fibrosis pulmonar** está asociado solo con la variable síntomas **hemoptosis**.
- La variable antecedente **bronquiestasis** está asociado con las variables síntomas **dísnea, tos crónica y hemoptosis**.

Desde el punto de vista médico, la asociación entre las variables antecedentes y síntomas son muy razonables y mejoran el panorama de análisis del trabajo original de De los Rios y Bravo (2012). Pero, el análisis estadístico bajo el principio de los modelos lineales generalizados, con la técnica del los modelos log-lineales son contundentes, los mismos que los presentamos en los siguientes cuadros:

Número de episodios de TBC:

	disnea	sibilancia	hemoptsis	nepisaño
chi.cuad	12.517	6.686	17.706	13.561
p-valor	0.051	0.035	0.001	0.035
trv	13.246	8.45	20.703	11.16
gl	6	2	4	6
u	1.4476	1.9667	1.7596	0.9943
u_i^A	0.9510	1.0108	1.0118	1.0118
	0.2765	0.3373	0.3373	0.3373
	-1.2275	-1.3491	-1.3491	-1.3491
u_j^B	0.5234	0.8118	0.3116	1.1325
	0.6157	-0.8118	0.3672	1.1852
	-0.7134		-0.6788	-0.5325
	-0.4257			-1.7852

La variable antecedente *número de episodios de TBC* es importante en presencia de la categoría del primer episodio ($0.95 \leq u_1^A \leq 1.01$) con las cuatro variables síntomas. La presencia de la variable síntomas de secuela *dísnea, hemoptisis y número de episodios año* en el paciente, en la segunda categoría es más importante que en la primera categoría ($u_2^B > u_1^B$); La presencia de *sibilancia* en el paciente ($u_1^B = 0.8118$), es importante.

Diagnóstico de asma (c1, c2, c3), diagnóstico de fibrosis (c4):

Estimador:	disnea	sibilancia	nepisaño	hemoptsis
chi.cuad	17.754	6.014	13.734	5.855
p-valor	0.007	0.049	0.033	0.054
trv	19.952	7.377	15.491	7.471
gl	6	2	6	2
u	1.5325	2.081	1.1085	2.2219
u_i^A	0.7984	0.7984	0.7984	0.8559
	-1.1268	-1.1268	-1.1268	-0.8559
	0.3284	0.3284	0.3284	
u_j^B	0.4988	0.8118	1.1325	0.3116
	0.6239	-0.8118	1.1852	0.3672
	-0.7052		-0.5325	-0.6788
	-0.4175		-1.7852	

La variable antecedente *diagnóstico de asma* en el paciente es importante en presencia de la primera categoría ($0.79 \leq u_1^A \leq 0.86$) y en la tercera ($u_3 1^A = 0.3284$) con las tres variables síntomas. La presencia de las variables síntomas de secuela: *dísnea y número de episodios año* en el paciente, en la segunda categoría es más importante que en la primera categoría ($u_2^B > u_1^B$); La presencia de *sibilancia* en el paciente ($u_1^B = 0.8118$), es importante.

La presencia de la variable antecedente *diagnóstico de fibrosis pulmonar* es importante ($u_1^A = 0.8559$); La presencia de la variable *hemoptisis* en el paciente es importante

en la segunda categoría, seguida por la primera ($u_2^B > u_2^B$).

Diagnóstico de bronquiestasis:

Estimador:	disnea	toscr	hemoptsis
chi.cuad	18.124	19.479	20.423
p-valor	0.006	0.003	0.0004
trv	15.751	18.366	17.102
gl	6	6	4
u	1.0586	0.8352	1.3999
u_i^A	1.5601	1.5601	1.5601
	-1.9057	-1.9057	-1.9057
	0.3456	0.3456	0.3456
u_j^B	0.4988	1.2528	0.3116
	0.6239	0.1541	0.3672
	-0.7052	-0.1141	-0.6788
	-0.4175	-1.2928	

La variable antecedente *diagnóstico de bronquiestasis* es importante en la primera categoría ($u_1^A = 1.5601$) con las tres variables síntomas, seguida de la tercera categoría ($u_3^A = 0.3456$). La presencia de la variable síntomas de secuela *dísnea y hemoptisis* en el paciente, la segunda categoría es más importante que en la primera categoría ($u_2^B > u_1^B$); La presencia de *tos crónica* en el paciente es importante en la primera categoría ($u_1^B = 1.25288$).

5.5. Análisis log-lineal Poisson con tablas de tres vías

Si en nuestro análisis adicionamos las variables *datos generales del paciente*, entre ellos **sexo, edad y IMC**, observamos que el problema de la TBC está asociado con *sexo* y de alguna manera con el *IMC*, pero es independiente de la edad. Por tanto, dedicaremos nuestra atención a la variable *sexo*, en cuanto a la siguiente hipótesis:

H_0 : las variables sexo del paciente, antecedentes del paciente y síntomas de secuela son independientes

vs

H_1 : existe asociación entre las variables sexo del paciente, antecedentes del paciente y síntomas de secuela

Para probar la hipótesis usaremos el test chi-cuadrado de Pearson (chi.cuad) y el test de razón de verosimilitud (trv), cuyos resultados los presentamos en el siguiente cuadro:

SEXO vs:	disnea	toscr	sibilanc	hemoptsis	nepisano
ttratm					
nepisod	XXX			XXX	
dasma	XXX		XXX		XXX
dfibrosis				XXX	
broqtsis	XXX	XXX		XXX	

Los resultados del contraste, nos muestra que la inclusión de la variable *sexo* no modifica la relación entre las variables *antecedentes* y *secuelas*, excepto *número de episodios* y *sibilancia* que no son significativos con la variabe *sexo*.

Asimismo, debemos advertir que los modelos log-lineal para tablas de 3 vías para nuestros datos: filas, columnas y profundidad, se dispersan demasiado, dando lugar a ceros en las celdas o contadas de 5 o menos. Para un mejor análisis de tres vías, es necesario un mayor número de observaciones, mayores a 200, que en la práctica es difícil y oneroso, dado que no existe un sistema de colección de datos sobre este problema, pese a su importancia desde el punto de vista médico y social.

Los valores de las estadísticas de la chi-cuadrado (chi.cuad) y de la razón de verosimilitud (trv) los presentamos en el siguiente cuadro, donde además, se muestran los intervalos de confianza del p-valor obtenidos por el método bootstrap, los mismos que muestran una buena estimación, tanto para el test de la chi chadrado y

de la razón de máxima verosimilitud, que asintóticamente tienen distribución chi cuadrado.

sexo	gl	chi.cuad	p-val	pv-boot	rvs	p-val	pv-boot
nepis * disnea * sexo							
H	3	7.926	0.048	[0.027, 0.036]	9.835	0.020	[0.025, 0.034]
M	6	7.647	0.265	[0.255, 0.278]	7.670	0.263	[0.335, 0.360]
T	6	12.142	0.059	[0.051, 0.063]	12.772	0.047	[0.054, 0.066]
nepis * sibilancia * sexo							
H	1	2.751	0.097		4.186	0.041	
M	2	3.932	0.140	[0.142, 0.160]	4.934	0.085	[0.097, 0.112]
T	2	6.686	0.035	[0.041, 0.052]	8.450	0.015	[0.013, 0.020]
nepisod * hemopt * sexo							
H	2	17.198	0.000	[0.000, 0.000]	21.512	0.000	[0.000, 0.000]
M	4	10.553	0.032	[0.024, 0.033]	12.201	0.016	[0.021, 0.029]
T	4	17.706	0.001	[0.001, 0.003]	20.703	0.000	[0.000, 0.001]
nepisod * nepisaño * sexo							
H	2	5.931	0.052	[0.060, 0.072]	6.499	0.039	[0.064, 0.077]
M	6	7.092	0.312	[0.305, 0.329]	6.417	0.378	[0.440, 0.465]
T	6	13.561	0.035	[0.031, 0.040]	11.160	0.084	[0.088, 0.104]
dasma * disnea * sexo							
H	6	16.640	0.011	[0.005, 0.010]	21.360	0.002	[0.000, 0.002]
M	6	14.966	0.021	[0.019, 0.027]	14.653	0.023	[0.026, 0.035]
T	6	17.754	0.007	[0.004, 0.007]	19.952	0.003	[
dasma * sibilanc * sexo							
H	2	4.540	0.103	[0.135, 0.153]	6.361	0.042	[0.079, 0.094]
M	2	2.273	0.321	[0.286, 0.309]	2.887	0.236	[0.286, 0.309]
T	2	6.014	0.049	[0.043, 0.054]	7.377	0.025	[0.024, 0.033]

sexo	gl	chi.cuad	p-val	pv-boot	rvs	p-val	pv-boot
dasma * nepisaño * sexo							
H	4	5.785	0.216	[0.195, 0.215]	7.672	0.104	[0.154, 0.173]
M	6	11.087	0.086	[0.078, 0.093]	10.529	0.104	[0.108, 0.125]
T	6	13.734	0.033	[0.027, 0.036]	15.491	0.017	[0.014, 0.021]
dfibrosis * hemoptosis * sexo							
H	2	0.446	0.800	[1.000, 1.000]	0.788	0.674	[1.000, 1.000]
M	2	6.090	0.048	[0.046, 0.057]	7.480	0.024	[0.027, 0.036]
T	2	5.855	0.054	[0.046, 0.057]	7.471	0.024	[0.025, 0.033]
bronqtsis * disnea * sexo							
H	3	1.226	0.747	[0.907, 0.922]	1.790	0.617	[0.907, 0.922]
M	6	17.801	0.007	[0.008, 0.013]	14.441	0.025	[0.015, 0.022]
T	6	18.124	0.006	[0.009, 0.015]	15.751	0.015	[0.007, 0.011]
bronqtsis * toscr * sexo							
H	2	2.186	0.335	[0.372, 0.398]	2.160	0.340	[0.372, 0.398]
M	6	16.050	0.013	[0.016, 0.023]	16.217	0.013	[0.007l, 0.012]
T	6	19.479	0.003	[0.004, 0.008]	18.366	0.005	[0.002, 0.004]
bronqtsis * hemoptsis * sexo							
H	2	0.842	0.656	[1.000, 1.000]	1.282	0.526	[0.817, 0.836]
M	4	21.890	0.000	[0.000, 0.000]	19.408	0.001	[0.000, 0.001]
T	4	20.423	0.000	[0.000, 0.001]	17.102	0.002	[0.000, 0.001]

En el cuadro podemos observar que los intervalos de confianza bootstrap del p-valor de los test de bondad de ajuste chi-cuadrado de Pearson y el de razón de verosimilitud se obtuvieron con 10,000 muestras aleatorias independientes con repetición, los mismos que contienen el p-valor del total, pero las tablas marginales por sexo muestran la paradoja de Simpson, en el sentido que conjuntamente las tres variables son significativas al 5 % o menos, pero las tablas marginales, en especial del grupo de hombres, no son significativos, con un p-valor mayor al 10 %. El motivo de esta contradicción los hubiésemos podido describir con el análisis loglineal para tablas

de tres vías, pero, no ha sido posible debido al número de datos considerados en la investigación.

6. Conclusiones

Las variables categóricas se presentan con frecuencia en investigaciones relacionados con la opinión pública, a cerca de de posicionamiento de marcas, productos, preferencias de los consumidores y de los ciudadanos, percepciones de las personas sobre la calidad de los servicios que hacen uso, de los personajes del ambiente político, etc. En medicina, en las investigaciones observacionales con pacientes, ya sea prospectivas o retrospectivas, respecto a una determinada enfermedad, se generan una serie de variables categóricas.

Si observamos cómo se hace el análisis de dichas variables, por ejemplo, tal como se nos presenta en la televisión, se reducen a tablas de frecuencias simples; en muchas investigaciones socioeconómicas y médicas, se aplican de manera directa el test de bondad de ajuste de la chi-cuadrada de Pearson, sin importar que dicha estadística de test tiene distribución asintótica ($n \to \infty$), y por tanto su validez se da solo cuando el tamaño de la muestra es el adecuado o suficiente.

Una solución al problema, para el mejoramiento del análisis de datos parara variables categóricas, se recurre a los modelos lineales generalizados, que permite obtener un modelo lineal que describa la relación entre una serie de variables independientes o explicativas $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ con una variable respuesta categórica Y, a través del predictor lineal $g(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, como son la regresión logística, la regresión poisson, la regresión binomial negativa y los modelos log-lineal.

Al construir los intervalos de confianza para estimar los parámetros β_j del modelo lineal y las pruebas de hipótesis, nos enfrentamos con serios problemas relacionados con la distribución de los estimadores y el cálculo de los errores estándar. En estos casos, el uso de la técnica del bootstrap, tanto paramétrico como el no paramétrico,

es importante, toda vez que simplifica los procesos de cálculos engorrosos y provee estimadores consistentes del error estándar.

En nuestro país, el mayor problema que enfrenta un investigador es la falta de un sistema de recopilación de datos confiables y suficientes, dado que cada investigador tiene que buscarlos o construir su propia base de datos, que es muy onerosa, más si es del tipo observacional, bien prospectiva o retrospectiva.

Referencias

[1] Bickel, P. y Doksum, K. (1976) *Mathematical Statistics: Basic Ideas and Selected* Topics. Holden Day Inc.

[2] Bickel, P. y Doksum, K. (2002) *Mathematical Statistics: Basic Ideas and Selected Topics* Vol 1, 2da Edic. Prentice & Hall.

[3] Cox, D. y Hinkley, D. (1974) *Theoretical Statistics*. Chapman & Hall.

[4] Silvey, S. (1970) *Statistical Inference*. Chapman & Hall.

[5] Agresti, A. (2002) *Categorical Data Analysis*. 2da Edic. Wiley & Sons.

[6] Agresti, A. (2007) *An Introduction to Categorical Data Analysis*. 2da Edic. Wiley & Sons.

[7] McCullagh, P y Nelder, J. (1983) *Generalized Linear Models*. Chapman & Hall.

[8] Nelder, J. y Wedderburn, R. (1972) *Generalized Linear Models*. JRSS series A. Vol 135 part 3.

[9] Fienberg, S. y Rinaldo, A. (2012) *Maximum Likelihood Estimation in Log-Linear Models*. The Annals of Statistics. Vol 40, Núm 2.

[10] Christensen, R. (1997) *Log-Linear Models and Logistic Models*. 2da Edic. Springer.

[11] Cox, D. R. (1970) *Analysis of Binary Data*. Chapman & Hall.

[12] Fienberg, S. (2007) *The Analysis of Cross-Classifief Categorical Data*. 2 edic. Springer.

[13] Paula, G. (2013) *Modelos de Regressao com apoio computacional*. IME - USP.

[14] Efron, B. (1979) *Bootstrap Methods: Another Look at the Jacknife*. The Annals of Statistics. Vol. 7, No. 1.

[15] Efron, B y Gong, G. (1983) *A Leisurely Look at the Bootstrap, the Jacknife, and Cross-Validation*. The American Statistician. Vol. 37, No. 1.

[16] Efron, B. y Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman & Hall.

[17] Efron, B. (1993) *The Jacknife, the Bootstrap and other Resampling Plants*. SIAM.

[18] Efron, B. y Stein, C. (1981) *The Jacknife Estimate of Variance*. The Annals of Statistics. Vol. 9, No. 3.

[19] Crawley, M. (2007) *The R Book*. Wiley.

[20] Cressie, N. y Read, T. (1984) *Multinomial Goodness of Fit Test*. Journal of the Royal Statistical Sosciety. Series B, Vol. 46, No. 3.

[21] Cressie, N. y Read, T. (1988) *Goodness of Fit Statistics for Discrete Multivariate Data*. Springer.

[22] Winkler, B. (1996) *Bootstrapping Goodness of Fit Statistics en Loglinear Poisson Models*. Sonderforschungsbereich 386, Paper 53. http://epub.ub.uni-muenchen.de/

[23] De los Rios, J. y Bravo, Y. (2012) *Protocolo de Investigación Secuela de la Tuberculosis Pulmonar: Espectro Clínico del Problema*. Hospital María Auxiliadora.