Universidad de Costa Rica - Sede de Occidente
Revista Pensamiento Actual - Vol. 15 - No. 25, 2015
ISSN impreso: 1409-0112 / ISSN electrónico: 2215-3586 ● 77 - 84

Diagnóstico de modelos de clases latentes en tablas con frecuencias pequeñas o nulas

Diagnosis of latent class models on tables with small or no frequencies

Carlomagno Araya Alpízar[1]

Recibido: 3/11/2014 / Aprobado: 20/10/2015

Resumen

El artículo trata del problema que puede surgir en la aplicación de los Modelos de Clases Latentes, cuando se tienen conjuntos de datos binarios con frecuencias de patrones de repuestas nulos o pequeños. En este trabajo se analiza el efecto de sumar una pequeña constante a las frecuencias observadas de cada patrón de respuesta en las celdas de la tabla múltiple sobre los errores relativos de los estadísticos de bondad de ajuste y las probabilidades totales de clases. Se determina que cuanto más pequeña sea la constante sumada a las celdas mayor exactitud tienen los resultados, esto por medio de la simulación de datos. Se recomienda utilizar la constante $1/2^p = 1/R$, donde p es número de variables con respuesta binaria y R representa el total de patrones de respuestas.

Palabras claves: clases latentes, tablas escasas, bondad de ajuste, datos binarios.

Abstract

The article deals with the problem that can arise in the application of latent class models, when you have sets of binary data with frequency patterns with no or small answers. This paper analyzes the effect of adding a small constant to the observed frequencies of each cell response pattern from the multiple tables of the relative errors of the goodness of fit test and total probabilities of classes. It is determined, by means of simulation data, that the smaller the constant added to the cells, the greater the accuracy of the results. It is recommended to use the constant $1/2^p = 1/R$, where p is the number of variables with binary answers and R represents the total of response patterns.

Keywords: latent class, sparse data, goodness-of-fit, binary data.

Introducción

El modelo de clases latentes (MCL) es una técnica estadística que permite estudiar la existencia de una o diferentes variables latentes a partir de un conjunto de variables explicativas observadas y definir con base en sus clases una tipología de los individuos analizados. La técnica utiliza los estadísticos de bondad de ajuste (EBA), los cuales son un criterio utilizado para la selección de un MCL. Estos estadísticos tienen, desde ciertas condiciones, una distribución de probabilidad teórica χ^2. La suposición es válida en relación con en el Teorema Integral de De Moire-Laplace, si tanto las frecuencias observadas de los patrones y el tamaño de la muestra son grandes ($f_{x_h} \to \infty, n \to \infty$).

Las celdas con frecuencias muy pequeñas o cero tienen una contribución muy importante en los EBA, al provocar un aumento del error Tipo I. Por esa razón, los EBA están muy distorsionados a medida que el número de celdas de la tabla múltiple tienen frecuencias observadas pequeñas. En este sentido, el objetivo de estudio es determinar el efecto de sumar una pequeña constante a la frecuencia de los patrones de respuestas sobre los errores relativos de los estadísticos de bondad de ajuste y las probabilidades totales de clases en MCL cuando se tienen frecuencias pequeñas o nulas en los patrones de respuesta.

1 Doctor en Estadística de la Universidad de Salamanca. Magister en Estadística de la Universidad de Costa Rica. Profesor en la Universidad de Costa Rica, Sede de Occidente, San Ramón, Costa Rica. Correo electrónico: carlomagnocr@gmail.com

El trabajo se desarrolla en las siguientes secciones: en la primera etapa se presenta una revisión de la literatura sobre los modelos de clases latentes y los EBA; a continuación se presenta la metodología utilizada para la simulación de dato según el número de variables manifiestas y clases latentes, y seguidamente se analizan los resultados mediante gráficos de errores relativos para los estadísticos de bondad de ajuste y probabilidades totales. En la última sección se presenta las conclusiones sobre el efecto que produce sumar una pequeña constante a las frecuencias de los patrones de respuestas en tablas escasas.

Métodos

El análisis de clases latentes es una técnica estadística multivariada propuesta por Henry & Lazarsfeld (1968), permite estudiar la existencia de una o varias variables latentes por medio de un conjunto de variables manifiesta. Las relaciones de dependencia entre las variables manifiestas de una tabla de contingencia, generalmente están determinada por la existencia de una asociación entre cada una de ellas y otra variable no observable directamente, llamada *variable latente*.

En este sentido, se distinguen dos tipos de variables en el modelo de clases latentes. Las variables que pueden ser directamente observadas, v*ariables manifiestas* (o indicadoras) conforman un vector de p componentes $X' = (X_1, \cdots, X_p)$ Las variables latentes son representadas por Y se expresan mediante el vector $Y' = (Y_1, \cdots, Y_q)$, donde $q < p$. Las variables observadas y las latentes se consideran variables categóricas con dos o más categorías. Matemáticamente,el MCL se representa partiendo de una matriz que contiene los resultados de *p* variables categóricas directamente observadas, las cuales se llaman variables manifiestas y serán denotadas como X_j, que conforman un vector columna de *p* componentes sobre una muestra total de *n* individuos (Araya, 2010).

Un caso particular del MCL se tiene cuando las variables manifiestas son binarias, es decir, solamente se tienen dos niveles de respuesta (0,1), por lo que la matriz de datos estaría constituida por solamente ceros y unos. Entonces, sea X' un vector de *p* variables manifiestas binarias, las cuales forman una tabla de contingencia *p*-dimensional considerada como indicadora de una variable latente Y y que estas variables definen un MCL con *C* clases o categorías. Las probabilidades condicionales de respuesta de cada una de las variables manifiestas dentro de la clase latente c para $c = 1, \cdots, C; i = 1, \cdots, p;\ x_i = 0,1$ siguen una distribución Bernoulli, esto es:

$$\pi_{X_i/Y(c)}(x) = \pi_{ic}^{x_i}\left(1 - \pi_{ic}\right)^{1-x_i}$$

donde π_{ic} es la probabilidad condicional de obtener una respuesta positiva en la variable X_i para un individuo de la clase latente c. Así, el MCL se puede escribir como:

$$\pi_X(x) = \sum_{c=1}^{T}\pi_Y(c)\prod_{i=1}^{p}\pi_{ic}^{x_i}\left(1 - \pi_{ic}\right)^{1-x_i}$$

Los parámetros del modelo están sujetos a las siguientes restricciones:

1) $\sum_{c=1}^{C}\pi_Y(c) = 1$

2) $\sum_{x_i=1}^{I}\pi_{x_i/Y(c)}(x_i) = 1$

Las probabilidades condicionales (2) son comparables con las cargas o loadings del análisis factorial. La expresión en (1) implica que la población puede ser dividida en C clases latentes exhaustivas y exclusivas, por lo tanto, la probabilidad conjunta de las variables manifiestas se obtiene sumando sobre la dimensión latente.

Para realizar las estimaciones de los parámetros del modelo, es decir, para estimar las probabilidades de clase, conjuntas y condicionales se utilizan procedimientos iterativos basados en estimaciones de máxima verosimilitud. Los más

conocidos son el algoritmo de Newton-Raphson (Haberman (1979) y el algoritmo Esperanza-Maximización (EM) (Dempster et al., 1977).

Por otra parte, la calidad del ajuste de un MCL puede determinarse con la comparación de las frecuencias observadas para cada patrón de respuesta f_x con las frecuencias estimadas $\hat{f}_x$ mediante el contraste Chi-cuadrado de Pearson o la razón de verosimilitud G^2. Las frecuencias esperadas están dadas por la expresión:

$$\hat{f}_x = n \times \left[\sum_{c=1}^{C} \hat{\pi}_Y(c) \prod_{i=1}^{p} \hat{\pi}_{X_i/Y(c)}(x_i) \right]$$

Luego se calcula χ^2 y G^2 como,

$$\chi^2 = \sum_{k=1}^{R} \frac{\left(f_x - \hat{f}_x\right)^2}{\hat{f}_x} \qquad G^2 = 2\sum_{k=1}^{R} f_x log\left(\frac{f_x}{\hat{f}_x}\right)$$

Cuando se tienen frecuencias esperadas menores a 5, Read & Cressie (1988) presentan otra alternativa a los test anteriores. Ellos proponen la utilización de una versión generalizada del estadístico Chi-cuadrado. La forma general del estadístico Read-Cressie es,

$$CR(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{k=1}^{R} f_x \left[\left(\frac{f_x}{\hat{f}_x} \right)^{\lambda} - 1 \right] \quad \lambda \neq 0$$

Por otra parte, Freeman & Tukey (1950) introdujeron el estadístico de bondad de ajuste FT^2 para probar H_0, que es dado por la expresión,

$$FT^2 = 4\sum_{k=1}^{R} \left(\sqrt{f_x} - \sqrt{\hat{f}_x} \right)^2$$

El número de grados de libertad (gl) de los estadísticos de bondad de ajuste para la distribución χ^2 se obtiene a partir de la diferencia entre el número de celdas de la tabla múltiple menos el número de parámetros por estimar en el modelo, o de igual forma,

$$gl = \left(2^p - 1\right) - \left[T * p + (T - 1)\right]$$

donde:

p = Número de variables manifiestas binarias.

T = Número de clases latentes para la variables latente Y.

Lo anterior se cumple cuando las tablas múltiples no presentan el problema de frecuencias pequeñas en los patrones de respuestas. En caso contrario, si las tablas son poco ocupadas (o sparse data en inglés) el supuesto de la distribución teórica no se cumple y será necesario utilizar otros métodos con el fin de evaluar la bondad de ajuste del modelo.

Si algunas de las frecuencias esperadas ($\hat{f}_x$) son ceros estructurales o ceros aleatorios, no podrán ser estimados los parámetros, aunque realmente existan. Si una celda de una tabla múltiple contiene un cero estructural, el correspondiente parámetro no existe, mientras que si se trata de un cero aleatorio dicho parámetro sí que existe, pero no puede estimarse a partir del conjunto de datos observados (Araya, 2010). En este sentido, Clogg & Goodman(1984) mostraron que el número de grados de libertad pasaría a ser igual al número de celdas sin ceros, menos el número de parámetros estimables.

La alternativa es sumar una pequeña constante a la frecuencia observada de cada patrón de respuesta, tiene su origen en los Modelos log-lineales[2]. Goodman (1974) recomendó utilizar este procedimiento cuando se tienen frecuencias bajas o nulas, mediante la suma de $\delta = 1/2$ a

2 El Modelo Log-Lineal es un método estadístico que tiene por objeto estudiar la clasificación de las variables categóricas o cualitativas. En esencia es un modelo de regresión lineal múltiple entre las variables categóricas y el logaritmo neperiano de la frecuencia de los patrones de respuesta, de la forma: $\log m_{X,Y}(x,c) = \mu + \lambda_{i_1}^{X_1} + \ldots + \lambda_{i_o}^{X_o} + \lambda_c^{Y} + \lambda_{i_1 c}^{X_1 Y} + \ldots + \lambda_{i_o c}^{X_o Y}$

la frecuencia de cada celda. Grizzle et al. (1969) y Johnson & Koch (1970) propusieron sumar a todas las frecuencias de la tabla $1/R$, donde es el número total de posibles patrones de respuestas. En este sentido, Agresti (1990) propone realizar un análisis de la sensibilidad con diferentes valores de δ, para medir el efecto sobre las estimaciones de los parámetros y sobre los estadísticos de bondad de ajuste (EBA). Esta alternativa resulta útil solo si existen pocas celdas con frecuencias bajas debido al incremento del tamaño de la muestra.

Metodología

Para medir el efecto de las constantes sobre los errores relativos de los EBA y las probabilidades totales de clases, fueron simulados conjuntos de datos de 1000 patrones, para cada nivel de densidad de patrones (1,2,…,20) y 5 constantes. En total, para cada MCL fueron generados 100,000 patrones. También, se consideró sumar la constante a todas las frecuencias de los patrones de respuesta o únicamente a aquellos con frecuencias nulas (celdas vacías), Es para variables manifiestas de 4 hasta 10 y tamaños de muestra $n = 2^p$ hasta $n = 20 \times 2^p$.

Se debe señalar que las constantes sumadas a la frecuencias de los patrones de respuestas y son las mismas para todas las simulaciones que se presentarán en el artículo, las cuales son: 0, R, 0.25, 0.50 y 0.75. Donde $R=1/2^p$ y 2^p representa el número de patrones de respuesta posible para *p* variables manifiestas binarias.

Resultados

En cuanto a los resultados, los EBA presentan diferencias promedios significativas en los errores relativos según las constantes sumadas a la frecuencia de las celdas (las celdas representan los patrones posibles) y los niveles de densidad de

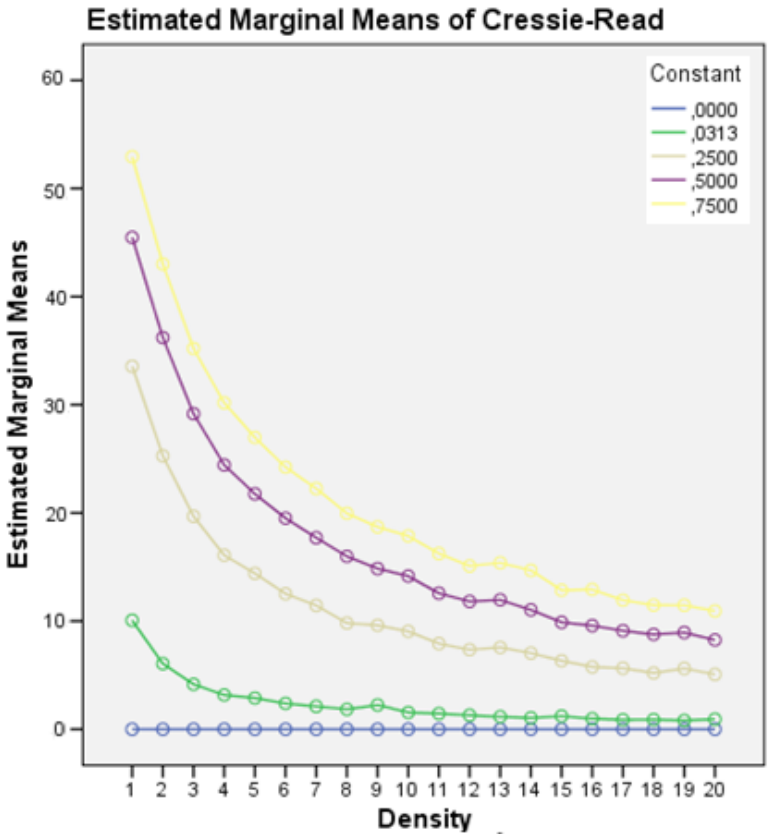

Gráfico 1. Errores relativos para el estadístico Cressie-Read: 5 variables manifiestas y 2 clases latentes (sumando constantes a todas las celdas)

En cuanto a las constantes, se encontraron diferencias altamente significativas en las medias de

los patrones, tanto para la alternativa de sumar la constante a todas las celdas o solamente a las vacías. En general, para todas las simulaciones existe una interacción altamente significativa entre la densidad de los patrones y las constantes, para las medias de los errores relativos de los EBA y probabilidades totales de clase.

Los errores relativos de los EBA son grandes cuando el nivel de densidad de los patrones es pequeño y disminuyen considerablemente al incrementarse el tamaño de la muestra y, por ende, al aumentar el porcentaje de patrones de respuestas observados, estos son aquellos que tienen frecuencias no nulas. Lo anterior se muestra en el **Gráfico 1**, por ejemplo para 5 variables manifiestas binarias y 2 clases latentes al sumar las constantes a todas las celdas y para el estadístico Cressie-Read.

los errores; por ejemplo, para el caso de 6 variables manifiestas, cuatro clases, sumando las constantes únicamente a las celdas vacías y un nivel de densidad de 5, el error relativo promedio para el estadístico de Pearson, usando la constante R es 9,34%, en tanto para constante 0,5 el error relativo medio es 40,29%. En general, la constante R presentan las medias de los errores relativos más pequeña en comparación a 0.25, 0.50 y 0.75.

Como se mencionó, los errores relativos decrecen al aumentar el nivel de densidad de los patrones, pero el efecto sobre los resultados de los errores relativos al usar R es más significativo, ya que tienden a ser muy pequeños y no presentar diferencias significativas respecto al grupo control, que representa los EBA calculados sin sumar ninguna constante a las frecuencias de los patrones de respuesta.

Cuando se tienen 8 variables manifiestas, 3 clases latentes y si se suman las constantes a todas

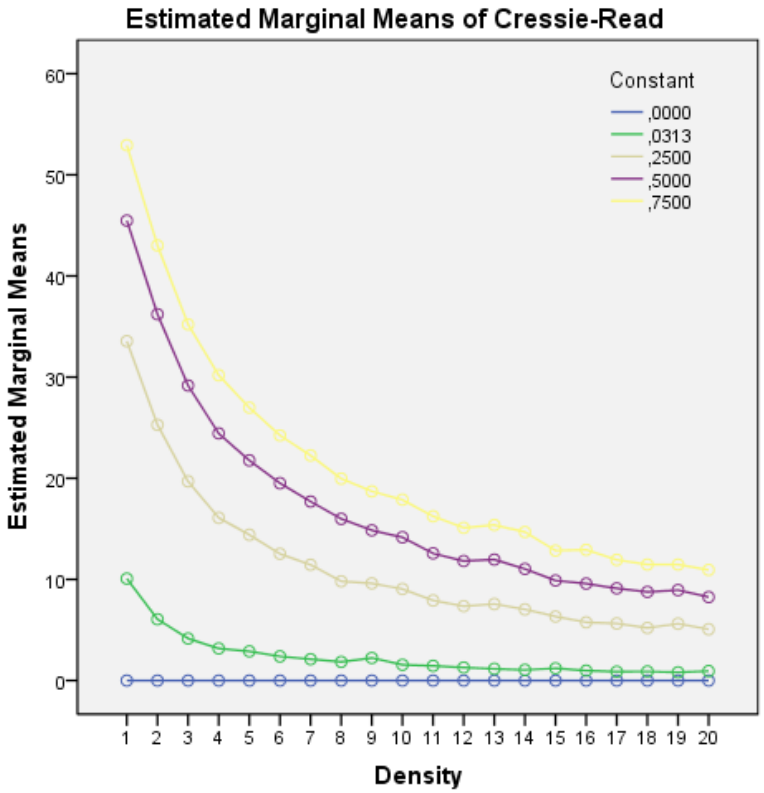

Gráfico 2. Errores relativos para el estadístico Cressie-Read: 8 variables manifiestas, 3 clases latentes y sumando constante a todas las celdas.

Para las otras constantes (0.25, 0.50 y 0.75), aunque los errores decrecen conforme aumenta la densidad

las celdas, el estadístico Cressie-Read para una densidad de 1, se indica que el error relativo medio es 10.1%, en tanto, cuando la densidad es 20, el error relativo medio disminuye significativamente a tanto solamente de 0,92% (**Gráfico 2**).

de los patrones, continúan presentando diferencias significativas entre ellos y fundamentalmente, en relación con la constante 0 y R.

El estadístico Freeman-Tukey presenta los errores relativos promedios más grandes, al tomar valores superiores al 75%. Por ejemplo, para 6 variables manifiestas, 3 clases latentes y la suma de la constante a las celdas vacías; para una densidad de 1, usando las constantes 0.50 y 0.75, los errores relativos promedios son 80,6% y 86,1%, respectivamente.

Esta situación que afecta al estadístico Freeman-Tukey se cumple para todos los niveles de densidad y constantes (**Gráfico 3**). Los estadísticos de Pearson y Cressie-Read asumen los errores relativos más pequeños conforme aumenta la densidad de patrones.

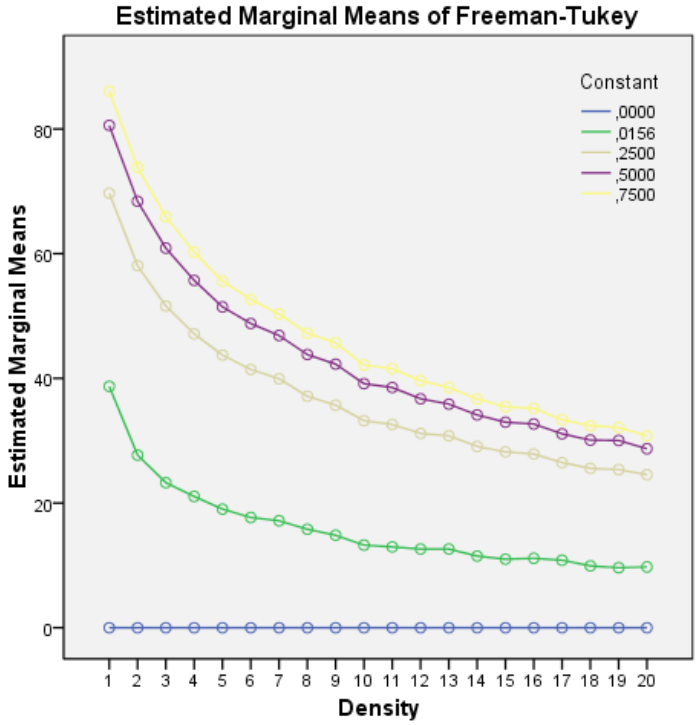

Gráfico 3. Errores relativos para el estadístico Freeman-Tukey:
6 variables manifiestas binarias y 3 clases latentes (con la suma de la constante a las celdas vacías)

Resulta importante mencionar que para el estadístico Cressie-Read, cuando las densidades son aproximadamente superiores a 10, las diferencias promedios entre los errores relativos de las constantes 0 y R tienden a ser mínimas y no significativas. Por lo general, en promedio los errores relativos son menores al 1,5% (**Gráfico 2**).

Por otra parte, respecto a los errores relativos sobre las probabilidades totales de clases del MCL, en general, los errores tienden a ser pequeños cuando los tamaños de las clases son grandes, y aumentan conforme las probabilidades totales son pequeñas, lo cual se observa en el **Gráfico 4** para clase latente 1, para 6 variables manifiestas y 2 clases latentes.

Al utilizar la constante 0.75 se obtienen los mayores valores medios de los errores relativos

para las dos clases latentes. Lo anterior, se cumple para la opción de sumar la constante únicamente a las frecuencias de patrones no observados en la muestra (es decir, celdas vacías), como también para opción de sumar la constante a todas las celdas.

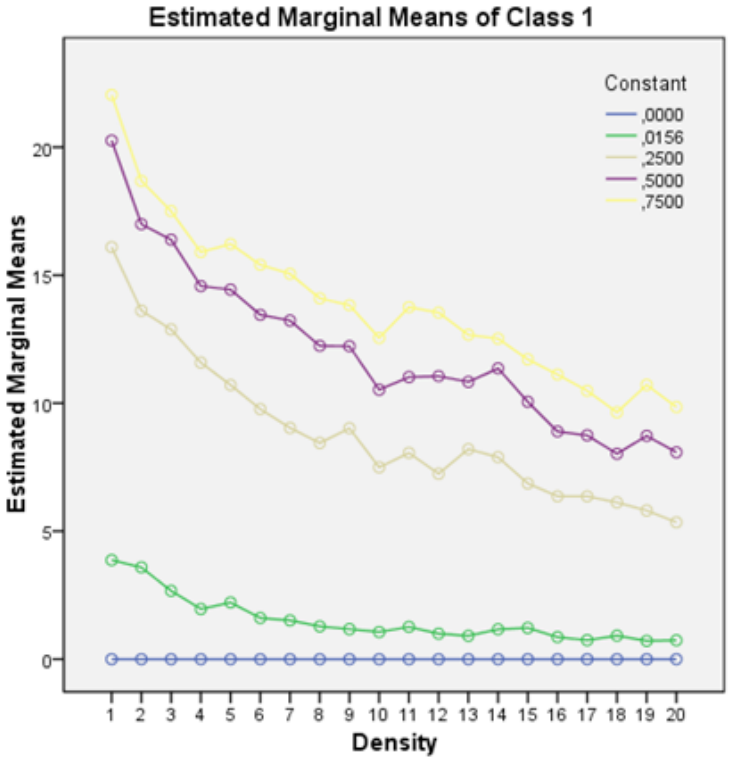

Gráfico 4. Errores relativos para la clase latente 1:
6 variables manifiestas y 2 clases latentes (sumando constante celdas vacías)

Conclusiones

Los resultados obtenidos permiten llegar a las siguientes conclusiones:

- El sumar una constante a las frecuencias de los patrones de respuesta (celdas) modifica significativamente los errores relativos de los estadísticos de bondad de ajuste y las probabilidades totales de clase.
- Se obtienen resultados similares en cuanto a los errores relativos, al sumar la constante a todas las celdas o únicamente a las vacías (es decir, aquellos patrones de respuesta con frecuencias nulas).
- El efecto los errores relativos al sumar la constante R, es más pequeño en comparación al obtenidos al sumar 0.25, 0.50 y 0.75.
- Entre los estadísticos de bondad de ajuste, el estadísticos Freeman-Tukey proporciona los errores relativos más grandes, en comparación a la razón de verosimilitud, Pearson y Cressie-Read.
- Y finalmente, los errores relativos son las probabilidades a priori (tamaños de las clases latentes), aumenta conforme el tamaño de la muestra decrece. Además, se obtienen los errores relativos más pequeños usando la constante R.

Bibliografía

Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

Araya, A.C. (2010). *Modelos de clases latentes en tablas poco ocupadas: una contribución basada en bootstrap* (Tesis de Doctorado). Universidad de Salamanca, España.

Clogg, C.C. & Goodman, L. A. (1984). *Latent Structure Analysis of a set of Multidimensional Contingency Tables*. Journal of the American Statistical Association, 79 (388), 762-771.

Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). *Maximum likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, Series B (Methodological), 39(1), 1-38.

Freeman, M.F. & Tukey, J.W. (1950). *Transformations Related to the Angular and the Square Root.* Annals Mathematical Statistics, 21(4), 607-611.

Goodman, L.A. (1974). *Exploratory Latent Structure Analysis using both Identifiable and Unidentifiable Models.* Biometrika, 61(2), 215-231.

Grizzle, J.E.; Starmer, C.F.; Koch, G.G. (1969). Analysis of Categorical Data by Linear Models. *Biometrics*, 25(3), 489-502.

Haberman, S.J. (1979). *Analysis of Qualitative Data: New developments,* Vol. 2. Academic Press, New York.

Henry, N.W. & Lazarsfeld, P. F. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Johnson, W.D. & Koch, G.G. (1970). *Analysis of Qualitative Data: Linear Functions.* Health Services Research, 5(4), 358-369.

Read, D, T. & Cressie, N. (1988*). Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer, New York.