Revista de Matemática: Teoría y Aplicaciones 1994 **1**(1) : 41–46
cimpa – ucr – ccss issn: 1409-2433

ANÁLISIS ESTADÍSTICO DE DATOS CATEGÓRICOS ORDENADOS

Breda Muñoz*

Resumen

Se propone un modelo para datos de panel categóricos ordenados utilizando el supuesto de un proceso estocástico subyacente de Markov, continuo en el tiempo y de espacio discreto. La estimación de la matriz de transición tridiagonal estabiliza la implementación del método de máxima verosimilitud. La estimación de las tasas de intensidad es muy natural.

Abstract

Ordered categorical panel data is modeled assuming an underlying continuous-time finite-space Markov chain. The tridiagonal form of the intensity matrix renders the implementation of the maximum likelihood method more stable numerically. The intensities describe rates of transitions between adjacent categories producing a very natural interpretation.

1. Introducción

Los análisis de mediciones repetidas de una variable categórica pueden clasificarse en

1. modelos de transición los cuáles modelan las probabilidades de transición.
2. modelos marginales, los cuáles modelan las probabilidades marginales.

En el caso del modelaje de las probabilidades de transición se han propuesto diferentes modelos. Anderson y Goodman (1856) introdujeron el uso de las cadenas homogéneas de Markov de orden mayor o igual a uno. Un inconveniente que presentaron estos modelos es la estimación de un sinnúmero de parámetros. Por ejemplo en el caso de una Cadena de Markov de primer orden con K categorías, es necesario estimar $K(K-1)$ parámetros. Obsérvese que conforme aumenta el número de categorías aumenta el número de parámetros

*Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica.

a estimar. Ademáss la incorporación de covariables al modelo no resultaba en una forma accesible.

Otro intento para modelar las probabilidades de transición se atribuye a Göttheen y Pruscha (1992). Adaptando un método sugerido por McCullagh (1980), modelaron las probabilidades de transición acumuladas e introdujeron covariables en el análisis. La implementación de este tipo de modelo requiere la especificación de una matriz de probabilidades de transición, la cual es usualmente especificada de manera subjetiva por el investigador.

Kalbfleish y Lawless (1985) analizaron un panel de datos categóricos asumiendo un proceso de Markov continuo pero con un espacio de respuesta discreto. La teoría de procesos estocásticos señala que la matriz de intensidades de transición determina la matriz de probabilidades de transición. Estos autores señalaron que en algunos casos la matriz de transición tiene una estructura muy simple por lo cual su parametrización requiere un número pequeño de parámetros. Además la incorporación de covariables es posible vía una función de enlace determinada por el investigador. Este modelo no considera, sin embargo, datos categóricos ordenados.

El modelo que a continuación se presenta, se basa en el que propusieron Kalbfleish y Lawless (1985). En él se asume, que existe una cadena de Marko subyacente y continua en el tiempo, y que por lo tanto, las únicas transiciones directas posibles ocurren entre categorías adyacentes. Esto determina una matriz de intensidades tridiagonal, con K categoías, implicando que sólo $2(k-1)$ parámetros deben ser estimados. Se asume además que posteriormente a la aplicación de una transformación apropiada (por ejemplo, la transformación logarítmica), el vector de parámetros de intensidad resultante, es una función lineal de las covariables.

La ventaja de considerar una matriz de intensidades de forma tridiagonal, radica no sólo en la reducción del número de parámetros, sino también en la implementación de los estimadores de Máxima Verosimilitud (MLE), los cuales resultan más estables que en el caso de matrices de intensidad de formas más generales. Otra ventaja del modelo propuesto, es el hecho que las intensidades estimadas describen las tasas de transición entre categorías adyacentes y estas últimas tienen una interpretación bastante natural.

Se pretende modelar y analizar cierto tipo de variables obtenidas de experimentos naturales o controlados, las cuales son medidas en forma repetida en el tiempo, y toman valores de un espacio de respuestas categórico, finito, y ordenado. Sin pérdida de generalidad, asuma que las categorías se denotan por $1, \cdots, K$ y que los sujetos son muestreados en las unidades de tiempo $i, \cdots, T$.

2. El modelo

Para cada individuo en la muestra, la variable respuesta $Y(t)$, conforma una serie de tiempo categórica. Generalmente también se registra un conjunto de covariables para cada individuo. En la práctica, las covariables se pueden clasificar en

1. estáticas y específicas del sujeto, como sexo, estado civil e indicadores de pertenencia a clases o grupos sociales.

2. dinámicas y comunes a todos los sujetos, como las variables utilizadas al modelar tendencias, efectos periódicos y efectos de intervención en series de tiempo.

Usualmente, las covariables de carácter sujeto específicas son variables discretas, y por lo tanto estratifican la población de interés. Dentro de cada estrato la población es homogénea y por ende expuesta a los mismos riesgos.

Como se ha señlado previamente, la variable respuesta es observada en un conjunto finito de unidades temporales, sin embargo es posible considerarla como un proceso continuo en el tiempo.

El ordenamiento de las categorías restringe la transición de la variable respuesta a categorías adyacentes. De esta forma, en la categoría más baja (más alta), la transición únicamente ocurre hacia la derecha (la izquierda). Sin embargo, es posible que la variable respuesta en dos unidades de tiempo consecutivas sea observada en categorías no adyacentes, esto es explicable por el hecho del supuesto de un proceso estocástico continuo en el tiempo, el cual permite la posibilidad de transiciones no observadas en medio de dos puntos muestrales.

La mayoría de los datos disponibles son presentados en forma agregada, como por ejemplo mostrando las frecuencias de las transiciones. Esto justifica, por lo tanto, la importancia de determinar las probabilidades de transición entre categorías.

Asumamos en el desarrollo de la teoría que sustenta el modelo, que las probabilidades son condicionadas en las covariables y que éstas son constantes entre dos instantes consecutivos del tiempo.

Para facilitar la notación se omitirán las covariables.

Defina

$$\begin{aligned} p_{uv}(s,t) &= Pr(\text{la respuesta es observa en la categoría } v \text{ en el instante t} \\ &\quad \text{dado que fue observada en la categoría } u \text{ en el instante s}) \\ &= Pr(Y(t) = v | Y(s) = u) \end{aligned}$$

Las probabilidades de transición son determinadas por las tasas de intensidad de transición en intervalos muy cortos de tiempo:

$$\begin{aligned} p_{uv}(t, t+\Delta t) &= q_{uv}\Delta t + o(\Delta t) \qquad u \neq v \quad (3{,}1) \\ p_{uu}(t, t+\Delta t) &= 1 + q_{uu}\Delta t + o(\Delta t) \qquad (3{,}2) \end{aligned}$$

donde q_{uv} $u, v \in \{1, \cdots, K\}, u \neq v$ son constantes no negativas Entre más alto sea el valor de q_{uv}, más alta es la probabilidad de transición de la categoría u a la categoría v, en un período corto de tiempo. Asumamos que las q_{uv} son constantes. En la prática las q_{uv} dependen de las covariables y son por lo tanto constantes a trozos. Es posible demostrar que

$$\begin{aligned} & q_{uu} + \textstyle\sum_{u \neq v} q_{uv} = 0 \\ & \Rightarrow q_{uu} = -\textstyle\sum_{u \neq v} q_{uv}. \end{aligned}$$

Asumiendo que las ecuaciones (3.1) y (3.2) son válidas para todo t, y denotando:

$$p_{uv}(t) = Pr(Y(t) = v | Y(0) = u)$$

se obtiene:

$$p'_{uv}(t) = \sum_r p_{ur}(t) q_{rv}(t)$$

lo anterior expresado en notación matricial es:

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q} \qquad (3{,}3)$$

donde la entrada (u, v) de $\mathbf{P}$(t) es p_{uv} y la entrada (uv) de $\mathbf{Q}$(t) es q_{uv}. Esta ecuación es conocida como la ecuación de Chapman-Kolmogorov. La condición inicial es $\mathbf{P}(0) = \mathbf{I}$. Es posible demostrar sin dificultad lo siguiente:

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

donde se cumple que para cualquier matriz cuadrada $\mathbf{M}$ es válido:

$$e^{\mathbf{M}} = \sum_{n=0}^{\infty} \frac{\mathbf{M}^n}{n!}$$

Si $\mathbf{M}$ es diagonalizable, entonces $\mathbf{M} = \mathbf{H}\boldsymbol{\Delta}\mathbf{H}^{-1}$ donde $\boldsymbol{\Delta}$ es una matriz diagonal y $\mathbf{H}$ es la matriz con j-ésimo vector columna igual al vector propio correspondente al j-ésimo valor propio de $\mathbf{M}$, el cual es la entrada (j, j) de $\boldsymbol{\Delta}$. Entonces $e^{\mathbf{M}} = \mathbf{H}e^{\boldsymbol{\Delta}}\mathbf{H}^{-1}$, donde $e^{\boldsymbol{\Delta}}$ es una matriz diagonal con entrada (j, j) igual al exponencial de la entrada (j, j) de $\boldsymbol{\Delta}$. Se refiere a los lectores a Cox y Miller (1968) para una introducción de las cadenas de Markov continuas en el tiempo y pruebas de los resultados precedentes.

El ordenamiento de las categorías implica que la cadena continua de Markov solamente transita entre categorías adyacentes, resultando una matriz de intensidades de transición tridiagonal $\mathbf{Q}$

$$\mathbf{Q} = \begin{bmatrix} -q_{12} & q_{12} & 0 & \cdots & 0 & 0 & 0 \\ q_{21} & -q_{21} - q_{23} & q_{23} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q_{K-1,K-2} & -q_{K-1,K-2} - q_{K-1,K} & q_{K-1,K} \\ 0 & 0 & 0 & \cdots & 0 & q_{K,K-1} & -q_{K,K-1} \end{bmatrix}.$$

Defina

$$\mathbf{q} = (q_{12}, q_{23}, \cdots, q_{K-1,K}, q_{2,1}, q_{3,2}, \cdots, q_{K,K-1}).$$

Asuma una función de enlace h que satisface

$$h(\mathbf{q}) = \mathbf{X}(t)\theta.$$

donde la función h es evaluada en cada entrada del vector $\mathbf{q}$. Sea $\mathbf{X}(t)$ una matriz $2(K-1) \times p$ dimensional, de funciones de las covariables y θ un vector $p \times 1$ dimensional de parámetros desconocidos. Se usará para la presentación del modelo la transformación logarítmica.

Ecuación (3.3) y el supuesto de una cadena de Markov implican que la entrada (u, v) de $\mathbf{P}(t, t+1)$ satisface

$$\mathbf{P}(t, t+1) = e^{\mathbf{Q}} \quad (\mathbf{3{,}4})$$

donde $\mathbf{P}$(t,t+1) es igual a $p_{uv}(t, t+1)$. En la práctica la matriz $\mathbf{X}$ puede ser sujeto específica y constante a trozos, en vez de constante como se está asumiendo en este artículo. Para estos casos más generales, donde $\mathbf{Q}$ es reemplazada por $\mathbf{Q}(t, \theta)$, el valor que la matriz obtendría en el instante t, aún es válida la ecuación (3,4) . Observe que por simplicidad se ha omitido el índice que indica al sujeto.

La estructura tridiagonal de $\mathbf{Q}$ permite la diagonalización de (Horn E. y Johnson, 1985). Esto permite que la matriz $\mathbf{P}(t, t+1)$ sea fácilmente calculada. Observe que cuando los valores de $\mathbf{q}'_{uv}s$ son pequeños, $\mathbf{P} \approx \mathbf{I}+\mathbf{Q}$ donde $\mathbf{I}$ es la matriz identidad. Esta aproximaciόm permite obtener valores iniciales para la implementación del método MLE.

3. Implementación del MLE

Para el caso general de la matriz $\mathbf{Q}$, Kalbfleish y Lawless (1985) presentan una implementación del MLE via el método de aproximación de Fisher.

El procedimiento será bosquejado a continuación para el caso de covariables dinámicas comunes.

En este caso las matrices $\mathbf{P}(t, t+1)$ son las mismas para todos los individuos de la muestra. Suponga que una muestra aleatoria de tamaño N es observada en T instantes sucesivos en el tiempo.

Sea $n_{u,v}(t, t+1)$ el número de sujetos que están en la categoría u en el instante i y en la categoría v en el instante j. La función de verosimilitud condicionada en la distribución muestral observada en el instante 1 es:

$$\begin{array}{lll} L(\theta) & = \prod_{t=1}^{T-1} \sum_{u,v=1}^{K} p_{uv}(t, t+1)^{n_{uv}(t,t+1)} & \\ \Longrightarrow l(\theta) & = log(L(\theta)) = \sum_{t=1}^{T-1} \sum_{u,v=1}^{K} log(p_{uv}(t, t+1)) n_{uv}(t, t+1) & \\ \frac{\partial l(\theta)}{\partial \theta_j} & = \sum_{t=1}^{T-1} \sum_{u,v=1}^{K} \frac{n_{uv}(t,t+1)}{p_{uv}(t,t+1)} \frac{\partial p_{uv}(t,t+1)}{\partial \theta_j} & j = 1, \cdots, 2K \end{array}$$

Kalbfleisch y Lawless (1985) desarrollaron una expresión para $\frac{\partial p_{u,v}(t,t+1)}{\partial \theta_j}$ $j = 1, \cdots, 2K$ basada en la expansiόm en series de $e^{\mathbf{Q}t}$. La expresión para las derivadas parciales propuesta aquí está basada en la ecuación de Kolmogorov (ver (3.3))

$$\frac{\partial \mathbf{P}(t,t+1)}{\partial \theta_j} = \mathbf{P}(t, t+1)\mathbf{H}\mathbf{V}\mathbf{H}^{-1} \qquad (\mathbf{4},\mathbf{1})$$

donde

$$\mathbf{V} = \mathbf{G}\mathbf{H}^{-\mathbf{1}} \frac{\partial \mathbf{Q}(t, \theta)}{\partial \theta_i} \mathbf{H}$$

y $\mathbf{G}$ es la matriz con entrada (i, j) igual a

$$\left(\int_0^1 e^{(-\lambda_i + \lambda_j)t} dt \right)$$

donde los $\lambda'_i s$ son los valores propios de $\mathbf{Q}(t, t+1)$ y $\mathbf{H}$ es la matrz donde el i^{th} vector columna es el vector propio correspondiente al i^{th} valor propio de $\mathbf{Q}$. Es posible derivar la equación (4.1) a través de análisis rutinario por lo que su prueba se omite.

La implementación de las segundas derivadas parciales es de la siguiente forma :

$$\frac{\partial^2 l(\theta)}{\partial\theta_i\partial\theta_j} = \sum_{t=1}^{T}\sum_{u,v=1}^{K}\frac{n_{uv}(t,t+1)}{p_{uv}(t,t+1)}\frac{\partial^2 p_{uv}(t,t+1)}{\partial\theta_i\partial\theta_j} - \sum_{t=1}^{T}\sum_{u,v=1}^{K}\frac{n_{u,v}(t,t+1)}{p^2_{u,v}(t,t+1)}\frac{\partial p_{uv}(t,t+1)}{\partial\theta_i}\frac{\partial p_{uv}(t,t+1)}{\partial\theta_j}$$

Kalbfleisch y Lawless (1985) approximaron la matriz Hessiana por su valor esperado. Aquí la matriz hessiana es obtenida al omitir el primer término en la fórmula precedente. La prueba de la razón de verosimilitud y la del estadístico de Pearson para la bondad de ajuste tienen distribución asintótica Chi-cuadrado, con $SK(K-1)$-dimensión (θ) grados de libertad, donde S es el número de estratos , K es el número de categorías y θ es el vector de parámetros. El método MLE fue implementado usando la función ms del $Splus$.

Conclusión

La reducción del número de parámetros a estimar, y por ende, el aumento en los grados de libertad, es uno de los logros del modelo propuesto. Por otro lado, no es difícil probar la convergencia asintótica del estimador de máxima verosimilitud al verdadero parámetro poblacional; esto proporciona una forma de evaluar de las estimaciones obtenidas. La aplicación de este modelo a la estimación de las probabilidades de transición de variables categóricas ordenadas, requiere del conocimiento profundo del marco teórico en el cual se sustenta el modelo observado. Esta metodología es aplicable a tablas multidimensionales siempre que el supuesto de transición ascendente o descendente se mantenga. También puede ser utilizado para estimar tablas de movilidad.

Referencias

[1] Agresti, A. (1989) "A Survey of models for repeated ordered categorical response data", *Statistics in Medicine*, vol. 8, pp. 1209–1224.

[2] Anderson, T. & Goodman, L. (1956) "Statistical Inference about Markov Chains", *Annals of Mathematical Statistics* vol. 28, pp. 89-110.

[3] Becker, R.; Chambers, J. & Wilks, A. (1988) *The New S Language*, Wadsworth & Brooks \ Cole Advanced Books & Software.

[4] Bentley, D. & Cooke,K. (1973) *Linear Algebra with Differential Equations*. Holt, Rinehart and Winston, Inc.,

[5] Billinsgsley, P. (1961) "Statistical methods in Markov chains", *University of Chicago*.

[6] Chambers, J. & Hastie, T. *Statistical Models in S*. Wadsworth & Brooks \ Cole Advanced Books & Software.

[7] Cox, D. & Miller, H. (1968) *The of Stochastic Processes*. John Wiley & Sons Inc., N.Y.

[8] Fahrmeir, L. & Kaufmann, H.(1987) "Regression models for non-stationary categorical time series", *Journal of Time Series Analysis*, vol. 8,# 2, pp. 147-160.

[9] Goodman, L.(1962) "Statistical Methods for analyzing processes of change", *American Journal of Sociology*, vol. 68, pp. 57–78.

[10] Göttlein, A. & Pruscha, H. (1992) "Ordinal time series models with application to forest damage data", pp. 113-118.

[11] Hagenaars, J. (1990) *Categorical Longitudinal data*. Sage Publications, pp. 13-87,147-183.

[12] Horn & Johnson (1985), *Matrix Algebra*.

[13] Kalbfleisch, J.& Lawless, J. (1985) "The Analysis of Panel Data under a Markov assumption", *Journal of the American Statistical Association* vol. 80, #392, pp. 863-871.

[14] Kaufmann, H. (1987) "Regression models for nonstationary categorical time series: asymptotic estimation theory", *The Annals of Statistics*, vol. 15, #1, pp. 79-98.

[15] McCullagh, P. (1980) "Regression models for ordinal data", *Journal R. Statistical Society, series B*, vol. 42,# 2, pp. 109-142.

[16] Sweeting, T. (1980) "Uniform asymptotic normality of the maximum likelihood estimator", *The Annals of Statistics*, vol. 8, pp. 1375-1381.

[17] Ware, J.; Lipsitz, S. & Speizer, F. (1988) "Issues in the analysis of repeated categorical outcomes", *Statistics in Medicine*, vol. 7, pp. 95-107.