Estudio de enfermedades progresivas usando un modelo de Markov de estados múltiples

Juan Carlos Salazar Uribe, PhD*
René Iral Palomino, Esp estadística*

Resumen

Los factores de riesgo y su grado de asociación con una enfermedad progresiva, tal como la enfermedad de Alzheimer o el cáncer de hígado, pueden identificarse usando modelos epidemiológicos; algunos ejemplos de estos modelos incluyen los de regresión logística, Poisson, log-lineales, regresión lineal y mixtos. En las ciencias médicas, el uso de modelos que tengan en cuenta no solo los distintos estados de salud que un participante experimenta a través del tiempo sino también las características propias de cada uno de ellos (por ejemplo, edad, género, características genéticas, etc.) parece razonable y justificado. En este artículo se discute una metodología que permite estimar el efecto de covariables asociadas con una enfermedad cuando la progresión o regresión de dicha enfermedad puede ser idealizada por medio de un modelo de estados múltiples (multi-state model) con varios estados que a su vez permite tener en cuenta la asociación de las mediciones tomadas en un mismo participante a través del tiempo. El método expuesto, que se basa en la propiedad de Markov se ilustra con datos simulados acerca de la enfermedad de Alzheimer. Finalmente, se discuten los méritos y las limitaciones de este enfoque.[*Salazar JC, Iral R. Estudio de enfermedades progresivas usando un modelo de Markov de estados múltiples. MedUNAB 2005; 8:202-7*].

Palabras clave: Enfermedad de Alzheimer, marcadores genéticos, modelos de estados múltiples, datos longitudinales, dependencia de Markov.

Summary

Risk factors and their degree of association with a progressive disease, such as Alzheimerís disease or liver cancer, can be identified by using epidemiological models; some examples of these models include logistic and Poisson regression, log-linear, linear regression, and mixed models. Using models that take into account not only the different health status that a person could experience between visits but also his/her characteristics (i.e. age, gender, genetic traits, etc.) seems to be reasonable and justified. In this paper we discuss a methodology to estimate the effect of covariates that could be associated with a disease when its progression or regression can be idealized by means of a multi-state model that incorporates the longitudinal nature of data. This method is based on the Markov property and it is illustrated using simulated data about Alzheimerís disease. Finally, the merits and limitations of this method are discussed. [*Salazar JC, Iral R. Progresive diseases study using Markov´s multiple stage models. MedUNAB 2005; 8:202-7*].

Key words: Alzheimer´s disease, genetic markers, multiple stage models, longuitudinal data, Markov´s dependence.

* Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín, Colombia.

Correspondencia: Dr. Salazar, Universidad Nacional de Colombia, Sede Medellín, Calle 59A No 63-020, bloque 21, Escuela de Estadística, Autopista norte. A.A. 3840. E-mail: jcsalaza@unalmed.edu.co

Artículo recibido: 23 de septiembre de 2005; aceptado, 21 de noviembre de 2005.

Introducción

Los modelos de compartimientos han demostrado ser herramientas útiles en el estudio de la evolución de enfermedades progresivas tales como el sida, la enfermedad de Alzheimer o ciertas formas de cáncer.[1-5] Específicamente, Kay[1] propuso un modelo mediante el cual pudo evaluar el efecto del cambio en los niveles de las alfa-fetoproteínas (un marcador de cáncer de hígado) en el riesgo de muerte en pacientes con carcinoma hepatocelular. Por otro lado, el modelo propuesto por Frydman[2], fue implementado usando una base de datos de 262 hemofílicos que recibieron transfusiones de sangre contaminada con el virus del SIDA; la implementación de este modelo permitió identificar que aquellos sujetos que recibieron un mayor número de transfusiones desarrollaron síntomas de SIDA más rápidamente que aquellos que recibieron menos transfusiones. Joly y Commengues,[4] trabajando con los mismos datos que usó Frydman[2] estimaron que el riesgo relativo de desarrollar SIDA era mayor en el grupo que recibió más transfusiones (RR=2.2, intervalo de confianza del 95% de 1.16 a 4.24). Finalmente, Salazar et al[5] estimaron que personas con bajo nivel de educación eran más propensas a sufrir trastornos mentales que aquellas con un alto nivel de educación.

El principal propósito de este artículo es discutir una metodología que se basa en la propiedad de Markov,[6] de tal manera que el estado de salud futuro de un individuo se puede predecir con el conocimiento del estado en el que él o ella se encuentran en la actualidad, lo que permite estimar el efecto de las covariables que pudieran estar vinculadas con una enfermedad particular que se caracteriza por el cambio de un estado a otro (transición) cuando esta se monitorea a través del tiempo. Considere por ejemplo el siguiente modelo respecto a la gripe común que representa la dinámica de un participante a través de dos estados claramente definidos (figura 1). Otro modelo, que idealiza la progresión de la enfermedad de Alzheimer, se ilustra a continuación (figura 2).

El proceso de identificación de factores de riesgo y marcadores genéticos o biológicos asociados con la evolución de una enfermedad podría estar influenciado por el hecho de que la información acerca de la progresión o regresión de una enfermedad usualmente está disponible únicamente a intervalos en el tiempo (visitas), por lo que el momento cronológico en el que ocurre un cambio de un estado a otro usualmente se desconoce. Para ilustrar, suponga que en un estudio de enfermedad mental en adultos los investigadores deciden considerar únicamente cuatro posibles niveles de la condición, así: 1: sano, 2: incipiente, 3: intermedio, y 4: declarado. Bajo estas condiciones los pacientes usualmente proporcionan información en la forma que se ilustra en la tabla 1.

Tabla 1. Esquema de recolección de datos en el tiempo en dos pacientes con riesgo de enfermedad mental.

Número del paciente	Historia del paciente
1	Visita 1 (Mes=0, Estado mental=Sano) Visita 2 (Mes=4, Estado mental= Incipiente) Visita 3 (Mes=12, Estado mental= Incipiente) Visita 4 (Mes=22, Estado mental=Sano)
2	Visita 1 (Mes=0, Estado mental= Intermedio) Visita 2 (Mes=6, Estado mental= Declarado) Visita 3 (Mes=12, Estado mental=Declarado)

El paciente número uno fue visto inicialmente en el estado 1 (sano); luego, cuatro meses más tarde, se registró una transición al estado 2 (incipiente) y, nuevamente, al mes 12 se observó al participante en el estado 2; finalmente, al mes 22, se registró una transición de vuelta al estado (sano). Por otra parte, el paciente dos fue visto inicialmente en el estado 3 (intermedio) y seis meses mas tarde se observó en el estado 4 (declarado), donde permaneció ya que ese estado es irreversible. Usualmente, la historia de un paciente incorpora no solo la respuesta registrada de su estado mental y la fecha de visita sino también los valores de factores de riesgo de interés tales como edad, sexo, historia familiar de demencia, eventos adversos de salud importantes, puntajes obtenidos en pruebas de habilidad mental, etc.

Las llamadas funciones de intensidad de transición[6] permiten identificar el papel de las distintas características de un individuo en las diversas transiciones y a su vez pueden ser relacionadas con modelos que tengan en cuenta no solo la potencial naturaleza aleatoria de las respuestas

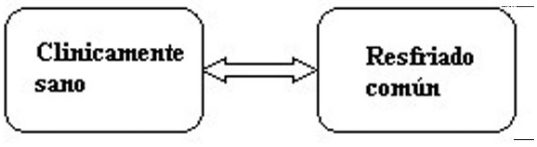

Figura 1. Modelo de dos estados que representa la dinámica del resfriado común. La flecha de doble entrada significa que una persona puede pasar de un estado a otro.

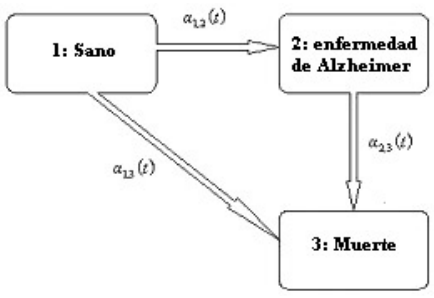

Figura 2. Modelo de tres estados con un estado absorbente (Muerte) La flecha en un solo sentido representa la naturaleza irreversible de los estados “enfermedad de Alzheimer” y “muerte”.

sino también el hecho de que las mediciones y por ende su variabilidad intrínseca se hacen a lo largo del tiempo. Por lo tanto, estudiar metodología que permita estimar estas funciones de intensidad de transición (o tasa instantánea de transición) tiene un valor importante e intrínseco desde el punto de vista de la investigación.[1]

Es pues nuestro interés discutir en este artículo cómo estas tasas instantáneas de transición que incorporan el efecto de las covariables (o factores de riesgo) pueden ser estimadas sin entrar en detalles y desarrollos técnicos que pudieran no ser de interés para una audiencia predominantemente no estadística.

El modelo

Se asume que la información acerca de la evolución de una enfermedad particular se puede representar por medio de un modelo de compartimientos (o de estados múltiples) como los ilustrados en la sección anterior, reconociendo que es posible definir otros modelos más complejos.[5] En este trabajo estamos interesados en ilustrar un modelo como el que se expone en la figura 2. Puesto que las observaciones que se registran en un sujeto en el tiempo están necesariamente correlacionadas (ya que provienen del mismo participante), se asume que dicha dependencia puede ser modelada asumiendo un proceso de Markov; esto quiere decir, que las transiciones futuras que un individuo pudiera tener, dependen de las transiciones pasadas únicamente a través de la última transición observada. La incorporación de esta dependencia de las observaciones intra-sujeto dentro del modelo debe mejorar la calidad de los estimadores de los efectos de las covariables.[10]

Bajo estos supuestos se usa un método iterativo que fue desarrollado con la finalidad de obtener estimadores confiables de las tasas instantáneas de transición[2, 5] corregidas por los distintos valores de los factores de riesgo de interés. Cabe mencionar que el vínculo entre los factores de riesgo y las funciones de intensidad de transición se puede establecer usando una parametrización que se asemeja en forma a un modelo de riesgos proporcionales de Cox;[9] sin embargo, otras maneras de calcular estos parámetros pueden ser utilizadas.[7]

Los estimadores de los efectos de las covariables se obtienen usando una función que se conoce con el nombre de función de verosimilitud y que está definida en la gran mayoría de textos básicos de estadística.[11] En vista de que este proceso iterativo de optimización de la función de verosimilitud requiere de la solución de un sistema de ecuaciones diferenciales (provenientes del supuesto de que el proceso evolutivo de la enfermedad obedece a un proceso de Markov) fue necesario desarrollar unas rutinas computacionales que se ejecutan usando el sistema SAS,[8] y que permiten obtener los estimadores de los efectos de las covariables para cada transición cuando se asume un modelo de tres estados como el ilustrado en la figura 2. Algunos detalles matemáticos con respecto al sistema de ecuaciones diferenciales y la función de verosimilitud se encuentran en el anexo. Estas rutinas están disponibles y pueden solicitarse a los autores.

Resultados

Con la finalidad de implementar e ilustrar el método se hizo uso de datos artificiales. Específicamente, se simuló una cohorte de 1.500 participantes. Después, para cada sujeto, se produjo una historia de transiciones basada en un máximo de seis visitas durante el periodo de seguimiento. Se asumió que a cada visita el participante podía ser clasificado en uno de tres estados mutuamente excluyentes que fueron definidos como: Sano, enfermedad de Alzheimer (declarado) y muerte; por supuesto, este último estado se registró en visitas diferentes a la primera. En la base de datos simulada, 442 sujetos con solamente la visita inicial fueron excluidos del análisis y solo se incluyeron aquellos con dos o más visitas ya que solamente estos proporcionan información de la evolución de la enfermedad bajo estudio a lo largo del tiempo (datos longitudinales). Por lo tanto un total de 1.058 participantes fueron incluidos en el análisis hipotético que aquí se informa.

Adicionalmente, y por simplicidad, se incluyó solo una covariable dentro del modelo: la edad del participante, pero cabe advertir que más covariables pudieron haber sido incorporadas en el modelo, incluyendo, posiblemente, interacciones entre dos o más covariables tales como educación, sexo, historia familiar de demencia, raza y eventos adversos serios como enfermedad cerebrovascular. Sin embargo, la inclusión de un alto número de covariables en este modelo particular podría comprometer la obtención de estimadores estables de los efectos de dichos factores debidos principalmente a razones de tipo computacional.

En esta base de datos se observó que los sujetos fueron vistos en promedio tres veces durante el periodo de seguimiento. La edad promedio al inicio del estudio fue de 75 (±4) años y de 84 (±6.7) años al final del estudio. El resumen de todas las transiciones registradas en la base de datos simulada (no de las personas que intervinieron en el estudio) aparece en la tabla 2.

Tabla 2. Número de transiciones que ocurrieron entre visitas para cada flecha en la figura 2.

Estado previo	Estado actual		
	Sano	Enfermedad de Alzheimer	Muerte
Sano	1255 (75.4%)	72 (4.3%)	337 (20.3%)
Enfermedad de Alzheimer	0	1192 (68.1%)	558 (31.9%)

La tabla 3 muestra los estimadores de los efectos de la covariable edad en las distintas transiciones. De acuerdo a los valores-P registrados en la misma tabla se observa que el efecto de la edad es estadísticamente significativo en cada una las transiciones de interés del modelo bajo consideración.

Tabla 3. Estimadores de máxima verosimilitud para el efecto de la edad en el modelo de tres estados.

Transición	Estimación	Error estándar	Valor-p
Sano->Alzheimer	0.028	(0.015)	0.0336
Sano->Muerte	0.057	(0.008)	<0.0001
Alzheimer->Muerte	0.059	(0.005)	<0.0001

La figura 3 ilustra el comportamiento de la tasa instantánea de transición para personas de 71, 81 y 91 años. Allí se nota que la probabilidad instantánea de transición se incrementa con la edad. Observe además que estas tasas de transición tienden a ser mayores en pacientes que padecen la enfermedad de Alzheimer. Por ejemplo, cuando se comparan pacientes de 71 años, la probabilidad instantánea de morir dado que se padece de la enfermedad de Alzheimer (0.048) es mayor que la probabilidad instantánea de morir dado que el sujeto está sano (0.022). Una tendencia similar se observa para otras edades (tabla 4).

Tabla 4. Tasas instantáneas de transición estimadas para el modelo de tres estados y para tres edades diferentes.

Edad	Transición		
	Sano a Alzheimer	*Sano a muerte*	*Alzheimer a muerte*
71 años	0.015	0.022	0.048
81 años	0.020	0.038	0.087
91 años	0.026	0.068	0.157

Discusión

En este trabajo se ha mostrado una forma de analizar datos que a nuestro conocimiento aún no ha sido explorada ni aplicada en Colombia y que promete ser una herramienta útil para el análisis de datos clínicos y epidemiológicos. A pesar de que la metodología se ilustra con datos artificiales es posible encontrar aplicaciones exitosas con datos reales.[1, 5, 7]

Entre las ventajas de este tipo de modelamiento es de destacar su flexibilidad para incorporar información que ha sido recopilada a través del tiempo, su interpretabilidad y su potencialidad para ser usado en situaciones donde se registren respuestas de tipo categórico; adicionalmente, se tiene el respaldo de que esta técnica esta basada en la función de verosimilitud, lo cual dota de propiedades es-

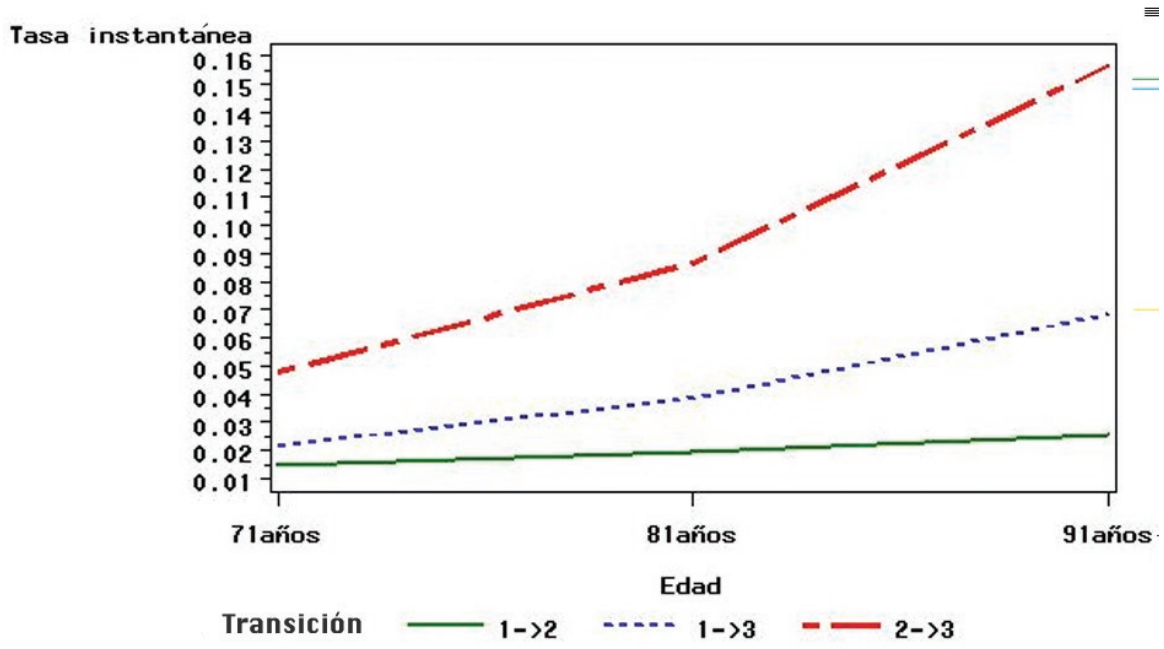

Figura 3. *Comportamiento de la tasa instantánea de transición durante un período de 20 años para el modelo de tres estados.*

peciales las estimaciones de los efectos de las covariables (por ejemplo, consistencia y normalidad asintótica que permite hacer inferencia).

Sin embargo, esta técnica requiere gran cantidad de datos, no incorpora información debida a censuras de intervalo,[12] no soporta un número muy elevado de covariables (algunos autores han reportado problemas trabajando incluso con un número reducido de variables en modelos complejos[4, 5, 13] y además, en muchos casos, es difícil justificar el supuesto de Markov. Sin embargo, al contrastar las ventajas y desventajas que ella posee, esta técnica es todavía una buena opción para analizar datos longitudinales con respuesta categórica.

Agradecimientos

Agradecemos a la Escuela de Estadística de la Universidad Nacional de Colombia, Sede Medellín por su apoyo para la realización de este trabajo. Un agradecimiento muy especial a los profesores Francisco Javier Díaz Ceballos y Juan Carlos Correa Morales por sus valiosos comentarios que mejoraron notablemente este manuscrito.

Referencias

1. Kay R. A Markov model for analyzing cancer markers and disease states in survival studies. Biometrics 1986; 42: 855-65.
2. Frydman H. Semiparametric estimation in a three-state duration dependent Markov model from interval-censored observations with application to AIDS data. Biometrics 1995; 51: 502-11.
3. Marshall G, Guo W, Jones RH. MARKOV: a computer program for multi-state Markov models with covariates. Comp Meth Progr Biomed 1995; 47:147-56.
4 Joly P, Commenges D. A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS. Biometrics 1999; 55: 887-90.
5. Salazar JC, Tyas SL, Snowdon DA, Desrosiers MF, Riley KP, Mendiondo MS, Kryscio RJ. Estimating intensity functions on multi-state Markov models with application to the Nun Study. Proceedings Joint Statistical Meeting San Francisco 2003; 3616-23.
6. Bhat UN. Elements of applied stochastic processes. New York: John Wiley and sons, 2 ed, 1984.
7. Harezlak J, Gao S, Hui SL. An illness-death stochastic model in the analysis of longitudinal dementia data. Statist Med 2003; 22:1465-75.
8. SAS Institute, Inc., SAS/IML Software: Usage and Reference, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1989; 501.
9. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer-Verlag, 2000: 39-44.
10. Aitkin M, Alfó M. Regression models for binary longitudinal responses. Stat Comput 1998; 8:289-307.
11. DeGroot M. Probabilidad y estadística. México: Addison-Wesley Iberoamericana, 2 ed, 1988: 302-4.
12. Meeker W, Escobar L. Statistical methods for reliability data. New York: Wiley Series in Probability and Statistics, 1998.
13. Faddy MJ. A note on the general time dependent stochastic compartmental model. Biometrics 1976; 32**:** 443-8.

Anexo. Ecuaciones diferenciales y función de verosimilitud

A partir del sistema de ecuaciones diferenciales de Kolmogorov

$$\frac{d}{dt}P(t) = P(t)\,Q$$

es posible encontrar una solución exacta para el modelo de tres estados ilustrado en la figura 2. Específicamente, se tiene que

$$\begin{bmatrix} p_{11}'(t) & p_{12}'(t) & p_{13}'(t) \\ 0 & p_{22}'(t) & p_{23}'(t) \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} p_{11}(t) & p_{12}(t) & p_{13}(t) \\ 0 & p_{22}(t) & p_{23}(t) \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -(\lambda_{12}+\lambda_{13}) & \lambda_{12} & \lambda_{13} \\ 0 & -\lambda_{23} & \lambda_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

Las ecuaciones a resolver son:

$$P_{11}'(t) = -(\lambda_{12}+\lambda_{13})P_{11}(t)$$
$$P_{12}'(t) = \lambda_{12}P_{11}(t) - \lambda_{23}P_{12}(t)$$
$$P_{13}'(t) = \lambda_{13}P_{11}(t) + \lambda_{23}P_{12}(t)$$
$$P_{22}'(t) = -\lambda_{23}P_{22}(t)$$
$$P_{23}'(t) = \lambda_{23}P_{22}(t) = -\lambda_{23}P_{23}(t)$$

Las soluciones usando factor integrante son, respectivamente:

$$P_{11}(t) = e^{-(\lambda_{12}+\lambda_{13})t}$$
$$P_{12}(t) = \frac{\lambda_{12}}{\lambda_{**}}\left[1-e^{-\lambda_{**}t}\right]e^{-\lambda_{23}t}$$
$$P_{13}(t) = 1 - P_{11}(t) - P_{12}(t)$$
$$P_{22}(t) = e^{-\lambda_{23}t}$$
$$P_{23}(t) = 1 - e^{-\lambda_{23}t}\ ,\ \lambda_{**} = \lambda_{12}+\lambda_{13}-\lambda_{23}$$

La contribución a la verosimilitud del k-ésimo individuo está dada por:

$$\prod_{i=1}^{M_k} P^{(k)}_{s_i, s_{i+1}}\left(t_i - t_{i-1}\right)$$

La función de verosimilitud para los *n* individuos está dada por:

$$L(\theta, X) = \prod_{k=1}^{n}\prod_{i=1}^{M_k} P^{(k)}_{S_i, S_{i+1}}\left(t_i - t_{i-1}\right) \quad , \quad \lambda_{ij} = f(\theta, X)$$