Un modelo de sobrevida multivariado para eventos recurrentes por sujeto con evento terminal: deserción de clientes en la industria de las Telecomunicaciones

Mariana Cárdenas Leuro
Estadística
Código: 832314

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Junio de 2013

Un modelo de sobrevida multivariado para eventos recurrentes por sujeto con evento terminal: deserción de clientes en la industria de las Telecomunicaciones

Mariana Cárdenas Leuro
Estadística
Código: 832314

Director
Luis Guillermo Díaz Monroy
Profesor asociado Universidad Nacional de Colombia

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Junio de 2013

Título en español

Un modelo de sobrevida multivariado para eventos recurrentes por sujeto con evento terminal: deserción de clientes en la industria de las Telecomunicaciones

Title in English

A multivariate survival model for recurrent events with a terminal event by subject : clients' desertion in Telecommunications industry

Resumen: El uso de modelos de sobrevida para estimar el riesgo de deserción de clientes en la industria de telecomunicaciones es común cuando los clientes son hogares o personas. En el segmento empresarial se observa un proceso subyacente asociado a la pérdida del cliente, no observado en los segmentos de hogares y personas, que es la desconexión de los servicios de manera paulatina. Esta situación no puede ser ignorada en el modelamiento del riesgo de deserción. En este trabajo se propone el modelamiento del riesgo de pérdida de clientes del segmento empresarial, en la industria de telecomunicaciones mediante un modelo de sobrevida multivariado para eventos recurrentes en presencia de un evento terminal.

Abstract: Using survival models to estimate the risk of clients' desertion in communications industry is frequent when clients are households or persons. In the enterprise segment an underlying process associated to client loss is observed, which does not appear in the segments of households and persons, and it is the slowpaced unplug of services. Such a situation can not be ignored in the modeling of desertion risk. The aim of this work is to propose the modeling of clients' loss risk for the enterprise segment, in telecommunications industry, through a multivariated survival model for recurrent events in the presence of a terminal event.

Palabras clave: Sobrevida, Datos de sobrevida multivariados, Modelos de fragilidad, Eventos recurrentes con evento terminal

Keywords: Survival, Multivariate Survival Data, Frailty models, Recurrent events with terminal event

Nota de aceptación

Trabajo de tesis

Jurado
Prof. Jaime Abel Huertas

Jurado
Prof. Luis Fernando Grajales

Jurado

Director
Luis Guillermo Díaz Monroy

Bogotá, D.C., Junio de 2013

Dedicado a

A Juan Felipe, mi hijo, ángel de mi vida y principal motor de mi existencia.
A mis sobrinos: Andrea, Daniela, Santi y Juanita. A quienes amo mucho.

Agradecimientos

Agradezco a Dios y a todos los ángeles que Él puso a mi lado para que me apoyaran, acompañaran y me dieran la fortaleza para hacer este posgrado y finalmente este trabajo:

A mi hijo y a mi esposo a quienes les debo todo, pues ellos fueron los que vivieron todo el proceso, permitieron y sacrificaron el tiempo que pudimos vivir juntos y que no disfrutamos por mis compromisos estudiantiles. Esto lo hicieron con admirable resignación. Muchas gracias por esto. Gracias cielito, gracias mi amor.

A mi madre por ser esa mujer hermosa, cálida y sencilla de quién heredé el ímpetu para hacer todas las cosas que me he propuesto en la vida y quien estuvo allí feliz creyendo en mí todo el tiempo. Gracias mamita.

A mi padre quien me dio el maravilloso consejo de estudiar esta magnífica carrera, él fue el origen.

A mi segunda madre, la Señora Elvia. Sin su ayuda y constante preocupación por mi hijo y por mí, yo hubiera claudicado.

A mis hermanos: Andrés, mi hermano también de vientre, quien estuvo en los últimos días (los de más cansancio) dándome todo su apoyo e incondicionalmente dispuesto a lo que necesitara. A mi hermana Tania, quien sabe perfectamente el significado de un abrazo; a Camilito, quien hizo las veces de madre ayudándole a mi hijo en sus tareas cuando yo no podía estar presente y a Alejito por motivarme con su fortaleza para asumir todas las cosas.

A mi director de tesis, el profesor Luis Guillermo Díaz Monroy, excelentísimo docente y persona, quien además de acompañarme y asesorame con paciencia y dedicación este trabajo, me compartió varios de sus sabios e inteligentes consejos, los cuales me sirvieron para sobrellevar situaciones difíles en tiempos difíciles.

A mi gran amigo Jonhatan quien me hizo valiosos aportes académicos y leyó mi documento esas veces en los que uno lo lee y no le vé los errores. Muchas Gracias Jhonatan.

A Holman García quien me dió el empujón necesario para continuar y terminar este posgrado en el momento indicado. Espero que desde el cielo estés contento como yo por este logro, amiguito Holman. Y finalmente a todos mis familiares no mencionados antes, pues sé que estuvieron muy pendientes todo el tiempo y me dieron voces de ánimo cuando más las necesitaba.

Un abrazo muy grande a todos. Dios los bendiga inmensamente.

Índice general

Índice de tablas

Índice de figuras

Introducción

En el área de mercadeo, para apoyar el cumplimiento de los objetivos estratégicos de las compañías, se invierte fuertemente en el control de la deserción de clientes pues ésta, además de impedir que se logren los objetivos de participación y penetración del mercado, es traducida inmediatamente en pérdida de los ingresos que, de pasar ciertos límites y dada la dinámica de competencia de los mercados, cuestan mucho recuperar. Las estrategias usadas para mantener clientes, además de que propenden por garantizar penetración de mercado, se convierten también en esquemas para generar una sólida relación cliente - empresa que favorece en última instancia la fidelización del cliente, estado de éste, bastante beneficioso para las compañías pues genera el codiciado aseguramiento de los ingresos. Cuando las tasas de fidelización de clientes son altas, los indicadores de valor de marca y reputación, aumentan en niveles nada despreciables, generando valor a la marca (Aaker & Biel (1993)). De esta manera, un adecuado control de la deserción redunda de manera inequívoca en el aumento de los índices de lealtad, de fidelización de clientes y valor de marca.

En el sector de las telecomunicaciones, se distinguen y se gestionan desde diferentes ópticas empresariales dos macrosegmentos desde el punto de vista de manejo de clientes: el masivo, compuesto por hogares y personas; y el empresarial. En éstos a su vez desde la óptica de producto se diferencian la telefonía fija y la telefonía movil. El segmento empresarial también se subsegmenta en otros dos diferenciables entre ellos por el tipo de productos que manejan y por el valor mismo de los clientes: el segmento de empresas grandes (segmento corporativo) y el segmento de las empresas no grandes (segmento pymes).

Aunque el segmento corporativo no es tan grande en unidades, comparado con los otros dos, es uno bastante rentable, ya que lo componen clientes que difícilmente pueden prescindir de servicios de telecomunicaciones. Esto facilita negociaciones bastante efectivas en rentabilidad. La pérdida de un cliente de este segmento afecta directamente el ingreso y puede afectar hasta la reputación de la compañía.

Los clientes corporativos acceden a una gran variedad de servicios de telefonía fija: Telefonía Local, Internet, Datos, Datacenter y servicios administrados entre otros. Estos servicios se conectan mediante *enlaces* que no son más que las conexiones que se hacen por las diferentes especificaciones del tipo de producto que el cliente solicita. Es frecuente que por la misma complejidad de la relación cliente - servicio, muchos de los clientes cuando toman la decisión de desertar, lo hagan empezando a desconectarse paulatinamente, lo que se reconoce como **deserción pasiva**. Por ejemplo, un cliente puede tener asociados 4 enlaces que pueden ser uno para un internet de 10 megas, otro para un internet de 4

megas y 2 enlaces de datos. La deserción pasiva de un cliente como este, puede empezar desconectando el enlace de 4 megas, luego continúa desconectando los dos de datos y finalmente el enlace de 10 megas. Lo que puede desencadenar finalmente en la pérdida total del cliente, y con toda seguridad, en la pérdia de los ingresos asociados a estos enlaces desconectados. En este sentido, se observa una recurrencia de la desconexión de los servicios hasta la desconexión total. Otros clientes, se desconectan en una sola ocasión desconectando todos los servicios.

Varias han sido las herramientas que desde el punto de vista estadístico se han desarrollado y aplicado con éxito para contar con argumentos técnicos que ayuden a entender, cuantificar y afrontar el problema de la deserción de clientes en la industria de las telecomunicaciones.

En el contexto de los modelos de sobrevida, el profesor Junxiang Lu en el 2002 defiende su posición de que aunque las herramientas estadísticas como los árboles de decisión y la regresión logistica son herramientas que han presentado resultados exitosos en términos de gestión de la deserción , la estimación de la probabilidad de sobrevida permite tener información predictiva del tiempo que duraría el cliente en las compañía lo que hace ser más efectivos en la implementación de estrategias para control de deserción. Además, en este mismo trabajo el profesor presenta una herramienta de clasificación de cliente partiendo del cálculo de la probabilidad de riesgo de pérdida por cliente. Este mecanismo de estimación toma gran relevancia porque con la estimación de la probabilidad de deserción del cliente o de sobrevivencia, se puede calcular el *valor del tiempo de vida del cliente* (**LTV**, Life Time Value)(Rosset et al. (2002), Lu (2003), Mutanen (2004)), indicador bastante útil desde el punto de vista de gestión financiera de los clientes ya que permite manejar el concepto de valor de cliente involucrando factores como la probabilidad de sobrevida del cliente, la rentabilidad del mismo y la tasa de descuento.

Por otro lado, desde la incursión de las herramientas de minería de datos en la industria, surgen mecanismos de clasificación de clientes basados en redes neuronales y árboles de decisión (Hung & Yen (2006)), herramientas que adoptaron el reto de estimación de la deserción de clientes mediante la clasificación de ellos en grupos de riesgo y no riesgo, tomando como variables de clasificación información transaccional y demográfica del cliente. Luego, se amplió el espectro de análisis a la realización de comparaciones entre los diferentes mecanismos estadísticos de clasificación existentes: los árboles de decisión , las redes neuronales, los modelos de regresión logistica y el análisis discriminate, en donde en general se encontró que las técnicas que mejor se desempeñaban eran la de árboles de decisión y la regresión logística(Tammadoni (2009), Kraljevic & Gotovac (2010) y Khalida et al. (2010)). Como lo indican varios autores que utilizan estas herramientas de análisis en sus estudios, efectivamente clasificar los clientes en propensos a desertar y no propensos, tiene el inconveniente que aunque se tiene un argumento estadístico muy sólido para, o bien gestionar con estrategias de retención a los propensos a desertar (otro producto por menos precio, meses gratis, etc), o bien gestionar con estrategias de fidelización a los no propensos a desertar (obsequios, invitaciones a eventos especiales, actualización de productos gratis, etc); para una proyección de presupuesto de ingresos este mecanismo es nada efectivo pues no permite predecir cuántos de estos clientes realmente continuarán con la compañía y cuántos realmente desertarán a un tiempo dado. Las cifras a los inversionistas normalmente contienen entre otras, estimaciones continuas de proyección de participación de mercado , por lo que, un análisis de sobrevida para este objetivo resulta ser una herramienta bastante útil.

Ahora, para el ajuste de los modelos de sobrevida mencionados, normalmente el tipo de modelamiento más utilizado es el del modelo de Cox en su versión clásica. Aunque también se encuentran algunos estudios con modelamientos de riesgos en competencia, que involucran la óptica del análisis de datos sobrevida multivariados (Alberts (2006), Braun & Schweidel (2011)), pero en muy poca proporción. No obstante, ninguno de los modelos mencionados, cubren el problema de ajustar la probabilidad de deserción de un cliente empresarial - corporativo, en los que se puede observar un proceso subyacente al riesgo de deserción de éste, debido a su historia de vida en cuanto a la tenencia y la desconexión de los servicios conectados. La particularidad de cómo se van desconectando servicios en este segmento, de alguna manera lo pone en condiciones diferentes de modelamiento con respecto a los demás. Igualmente en la bibliografía encontrada no se hace referencia a que la granularidad de los análisis incluya este segmento. En general, los modelos propuestos se refieren a modelamientos para la gestión de la deserción en segmentos masivos (hogares - personas) ya sea para telefonía móvil o fija. La gran atención en el segmento masivo se debe principalmente a que la distribución de ingresos del masivo vs el corporativo es aproximandamente de 75 % vs 25 % respectivamente; sin embargo la alta rentabilidad y economía de escala que se puede manejar con los clientes corporativos, hacen de este último un segmento muy valioso.

En este trabajo se adapta una estrategia de modelamiento para la estimación del riesgo de pérdida de los clientes del segmento corporativo, del negocio del telefonía fija, mediante el anális del tiempo de sobrevida de los clientes expuestos a eventos recurrentes con un evento terminal. Se toman las desconexiones de los enlaces como los eventos recurrentes. La recurrencia se refiere a la desconexión secuencial de los productos conectados (van desconectando los enlaces asociados a cada producto) y el evento terminal se homologa a la desconexión total o pérdida del cliente. Con la adaptación de este tipo de modelamiento se propone un nuevo enfoque para estimar la probabilidad de sobrevida del cliente incluyendo la historia del cliente en la compañía en cuanto a tenencia de productos. De esta manera, el objetivo de esta tesis es adaptar un modelo de sobrevida multivariado de eventos recurrentes por sujeto con evento terminal, para estimar el riesgo de pérdida de clientes del segmento de empresas grandes en una empresa de telecomunicaciones.

El desarrollo de este trabajo, se presenta en el siguiente orden: en el capítulo 1, se muestran los elementos teóricos que sustentan el modelamiento, en el capítulo 2, se presenta la metodología de modelamiento, en el capítulo 3 se encuentra la metodología de adaptación del modelo, la aplicación del modelo propuesto y finalmente en el último capítulo se presentan las conclusiones y recomendaciones.

CAPÍTULO 1

Elementos teóricos

1.1. La Función de sobrevida o de supervivencia

La sobrevida o tiempo de supervivencia se entiende como el período de tiempo desde el inicio de la observación hasta que un evento ocurre. Las funciones de sobrevida modelan el tiempo hasta la ocurrencia de un evento, entre otros, la muerte. En general el **tiempo** es una una variable de tipo contínuo [1] en cualquer unidad como mes, año, día, segundos, etc, medida desde el inicio de la observación del individuo en estudio hasta que el evento de interés ocurre. Por **evento** se entiende comúnmente, la muerte o la recaída de una enfermedad, sin embargo, la ocurrencia de cualquier hecho en el tiempo que sea objeto de análisis de sobrevida puede denotarse como evento: el nacimiento de un hijo, la ocurrencia de un accidente en una planta industrial, la adquisición vivienda propia, etc ; para el caso de este trabajo, el momento en el que se deja de ser cliente de una compañía por la desinstalación de los productos o servicios adquiridos. A la variable tiempo hasta el evento observado se le denomina **tiempo de sobrevida** o **supervivencia** y al evento **la falla**. Denominar al evento "falla "o "muerte "se debe principalmente a que los análisis de superviviencia se iniciaron en el contexto de áreas de la salud donde se estudian eventos relacionados generalmente con la muerte o deceso de los individuos estudiados en un período de tiempo, pero el evento puede ser definido y extendido según las condiciones y requerimientos del estudio.

Se define $S(t) = P(T > t)$ como la función de sobrevida o superviviencia de una variable aleatoria T cuyo recorrido toma valores en el intervalo $[0, \infty)$. $F(t) = P(T \leqslant t)$ es la función de distribución acumulada de la variable aleatoria T y la fuerza de mortalidad, función de intensidad o función de riesgo, la cual evalua el riesgo de muerte (falla) instantánea en el tiempo t condicionada al tiempo de sobrevida, se define como :

$$\lambda(t) = \lim_{\delta t \to 0} \Big[\frac{P(t < T \leqslant t + \delta t | T \geqslant t)}{\delta t} \Big] \tag{1.1}$$

[1] No obstante que el tiempo es una variable de tipo continuo, ésta puede considerarse o registrarse de forma discreta.

Se tienen las siguientes relaciones básicas entre la función de sobrevida, la función de riesgo y la función de distribución acumulativa de una variable aleatoria T, cuyas demostraciones se encuentran en Smith (2002):

- $S(t) = P(T > t) = \int_t^\infty f(u)du$, esto es : $S(t) = 1 - F(t)$
- Sea $f(t)$ la función de densidad de probabilidad de la variable aleatoria T, de tal manera que $F(t) = P(T \leqslant t) = \int_0^t f(u)du$, entonces $S(t) = Pr(T \geq t) = \int_t^\infty f(u)du$
- $$\lambda(t) = \frac{f(t)}{S(t)} \tag{1.2}$$

  Relación que se desprende de (1.1):

  $$\begin{aligned}\lambda(t) &= \lim_{\delta t \to 0}\Big[\frac{P\{t < T \leqslant t + \delta t | T \geqslant t\}}{\delta t}\Big] \\ &= \frac{F^{'}(t)}{S(t)} = \frac{f(t)}{S(t)},\end{aligned}$$

  que es la tasa instantánea de muerte en el tiempo t, dado que el individuo sobrevive hasta el tiempo t.
- de (1.2), la función de sobrevida se expresa en términos de la función de riesgo mediante la siguiente relación: $S(t) = e^{-\int_0^y \lambda(u)du}$

En análisis de datos de sobrevida una característica propia es la de censura de los datos. Se entiende por variable censura la indicadora que determina si el individuo experimentó o no el evento estudiado. Los datos censurados se representan en una estructura de datos mediante las variables de "censura". En las muestras observadas para los estudios de sobrevida sucede que al terminar el tiempo de estudio algunos individuos no experimenten el evento estudiado. Estos datos resultan ser censurados y esta censura se define como censura a derecha. Kleinbaum & Mitchel (2005) definen la censura cuando: "no se conoce exactamente el tiempo de supervivencia"para un individuo.

Los tipos de censura más comunes son a derecha: cuando se ha terminado el tiempo de estudio y no se observa el evento sobre un individuo, es decir tal vez le ocurrirá en un tiempo futuro; a izquierda: cuando al individuo le ha sucedido el evento antes del inicio del tiempo de estudio, y por intervalo: cuando sólo se sabe que al individuo le ha ocurrido el evento de interés dentro de un intervalo de tiempo, de tal maneraa que tales datos resultan ser "censurados por intervalo"

Los datos de sobrevida normalmente se presentan como la dupla (t_i, δ_i) donde t_i es el tiempo de observación y δ_i es una variable indicadora de censura, esto es, $\delta_i = 0$ si la observación es censurada y $\delta_i = 1$ cuando se observa la ocurrencia del evento de interés.

El tipo de censura o los tipos de censura que se observan en el estudio impactan directamente la función de verosimilitud, situación que afecta la estimación de los parámetros del modelo.

Es sabido que la presencia de datos censurados dificulta el escenario de que la función de sobrevida pueda ser obtenida directamente mediante métodos probabilísticos. Existen

varias maneras de estimar la función de sobrevida, entre los estimadores más conocidos y utilizados se encuentra el de Kaplan & Meier (1958). Este estimador tiene la gran ventaja de que se obtiene utilizando los mismos tiempos de observación, es decir surge de la tabla misma de datos de tiempo de sobrevida registrados en el estudio, no es necesario construir períodos de tiempo como sucede para otros tipos de estimadores propuestos. Una presentación de este estimador se puede encontrar en Klein & Moeschberger (1997) .

Ahora, en general, los estimadores disponibles para estimar la función de sobrevida así como el de Kaplan & Meier no tienen en cuenta variables explicativas que conduzcan a entender la relación entre la tasa de supervivencia y el tiempo mediante aquellas. Para esto, el modelo de Cox (1972) además de ser una herramienta eficaz para determinar estas posibles relaciones entre las variables (covariables) y el riesgo (también conocido como la tasa de mortalidad) de experimentar el evento de estudio a un momento dado, es la herramienta dispuesta desde la teoría de los análisis de regresión para el modelamiento de datos censurados.

1.1.1. El Modelo de Cox

El modelo de Cox dispone la manera de encontrar la relación entre la tasa de riesgo y algunas covariables asociadas con los individuos.

En el modelo de regresión de Cox, el riesgo para el i-ésimo individuo se define mediante la siguiente expresión:

$$\lambda_i(t) = \lambda_0(t) \exp\{Z_i\beta\} \tag{1.3}$$

donde $\lambda_0(t)$ es una función no negativa, Z es un vector de p covariables posiblemente tiempo dependientes para el i-ésimo individuo en el tiempo t y β es un vector fijo de p coeficientes de tamaño $p \times 1$.

En este modelo, se puede distinguir una parte paramétrica y una no-paramétrica: i) la parte paramétrica es: $Z_i\beta$, contiene el vector de parámetros de la regresión y se le denomina generalmente puntaje de riesgo (risk score) y ii) la parte no paramétrica: $\lambda_0(t)$, que es denominada función de riesgo base, es una función arbitraria y no especificada.

Al modelo de Cox se le denomina también de riesgos proporcionales por cumplir la siguiente propiedad:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp\{Z_i\beta\}}{\lambda_0(t) \exp\{Z_j\beta\}} = \frac{\exp\{Z_i\beta\}}{\exp\{Z_j\beta\}} \tag{1.4}$$

Esta característica del modelo supone que los riesgos para dos conjuntos diferentes de valores de las covariables, conservan la misma proporción en cada punto del tiempo.

En el modelo de Cox (1972) la censura debe ser independiente del valor futuro del riesgo del individuo, pues si se encuentra dependencia, la distribución de la variable respuesta podría ser seriamente sesgada, lo cual conduciría a obtener estimaciones sesgadas.

En los últimos años, las bases teóricas del modelo de Cox se han robustecido mediante el involucramiento de la teoría de las martingalas y los procesos de conteo. Es así que la propuesta del modelo de Cox ha tenido un replanteamiento teórico, situación que ha permitido utilizar una teoría más robusta del modelo de riesgos proporcionales de Cox

para el modelamiento de otro tipo de datos que se han venido presentando en los análisis de sobrevida. Es el caso de datos de sobrevida de tipo multivariado - cuando hay más de un evento (tiempo hasta un evento) registrado por individuo - . El constructo teórico presentado en esta tesis se basa en esta nueva formulación de la teoría para el análisis de datos de sobrevida. A continuación se presentan algunos elementos teóricos para las construcción de este modelo.

1.1.1.1. Procesos de conteo

Una variable aleatoria $N(t)$ representa un proceso de conteo sobre $[0, \infty)$ si:

1. $N(t)$ es un entero no-negativo, se asume que $N(0) = 0$
2. $N(s) \leqslant N(t)$ para $s < t$
3. $dN(t) = N(t) - N(t^-)$ es 0 o 1, donde $N(t^-)$ denota a $\lim_{\delta \to 0} N(t - \delta)$
4. $E(N(t)) < \infty$

El proceso de conteo $Ni = \{N_i(t) : t \geqslant 0\}$ cuenta el número de eventos de la unidad i, la ocurrencia se asume dentro del intervalo $(0, t]$. Un proceso de conteo puede ser considerado como el que registra y cuenta el número de eventos sobre una unidad en el tiempo de estudio.

Ahora, se define un proceso de conteo multivariado por:

$$N = (N_{ij} : i = 1, ..., n; j = 1, ..., m_i) \tag{1.5}$$

donde i indica la i-ésima unidad y j indica j-ésima ocurrencia del evento sobre el individuo $i = 1, 2, ..., n$. El proceso de conteo para el i-ésimo individuo es escrito por el vector $N_i = (N_{i1}, ..., N_{im_i})$.

1.1.1.2. El Modelo de Cox en el marco de los procesos de conteo

Para la introducción del modelo, desde el punto de vista de los procesos de conteo, se tiene que en el modelamiento clásico de sobrevida las principales variables incluidas son: la variable tiempo para evento T_i^*, el tiempo de censura C_i; con $T_i = min(T_i^*, C_i)$, la variable indicadora δ_i; la cual es igual a 1 si T_i^* es observada y 0 si la observación es censurada. Así, los datos consisten del par (T_i, δ_i). En la formulación del proceso de conteo, el par (T_i, δ_i) es reemplazado por el par $(N_i(t), Y_i(t))$ donde $N_i(t)$ es el número de eventos en el intervalo $[0, t]$ para la unidad i y $Y_i(t)$ es un indicador de riesgo en el tiempo t. ($Y_i(t) = 0$ si el individuo no ha experimentado el evento al tiempo t o $Y_i(t) = 1$ si al individo le ocurrió el evento al tiempo t).

Para la estimación de los parámetros β en (1.3) y con tiempos de sobrevida sin empates[2], bajo la estructura de los procesos de conteo, Andersen & Gill(1982) proponen la estimación de β basado en la función de verosimilitud parcial:

[2]Tiempos de sobrevida iguales para varios individuos

$$PL(\beta) = \prod_{i=1}^{n} \prod_{t \geqslant 0} \left[\frac{Y_i(t) r_i(\beta, t)}{\sum_j Y_j(t) r_j(\beta, t)} \right]^{dN_i(t)} \tag{1.6}$$

donde $Y_i(t)$ es la indicadora de que el individuo i aún se encuentra en observación al tiempo t, $N_i(t)$ es el número de fallas observadas para el sujeto i y $dN_i(t)$ es el incremento en $N_i(t)$ sobre el espacio de tiempo infinitesimal $[t, t + \delta t]$. $r_i(\beta, t)$ es el puntaje de riesgo para el sujeto i: $r_i(\beta, t) = \exp[Z_i \beta] \equiv r_i(t)$.

Realizando la diferenciación con respecto a β del logaritmo de la verosimilitud parcial se encuentra el vector de puntajes $U(\beta)$ de tamaño $p \times 1$:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} [Z_i(s) - \bar{z}(\beta, s)] dN_i(s) \tag{1.7}$$

donde $\bar{z}$ es la media ponderada de Z, sobre todas la observaciones que aún se encuentran en riesgo al tiempo s. β se obtiene solucionando la ecuación $U(\beta) = 0$. Este estimador es consistente y distribuido asintóticamente normal, con media β y varianza $I^{-1}(\hat{\beta})$, la inversa de la matriz de información observada.

Como se anotó anteriormente, si bien el planteamiento del modelo de Cox cubre la necesidad de llegar a modelar el tiempo hasta un evento de un grupo de sujetos, en Andersen & Gill (1982) se presenta este mismo modelo planteado desde el punto de vista de la teoría de martingalas la cual nace de la formulación del proceso de conteo de $\lambda(t; z) = \lambda_0(t) \exp(\beta_0'(t))$ $(t \geqslant 0)$.

Una de las ideas fundamentales de este desarrollo es que se demuestran las propiedades asintóticas de los coeficientes del modelo de Cox y además introduce una teoría que permite el desarrollo de modelos para el análisis de tiempos de sobrevida multivariados. El desarrollo supone que no puede ocurrir más de un evento al mismo tiempo.

En adelante se muestra la formulación del modelo de Cox, en el marco de los procesos de conteo multivariados según Andersen & Gill (1982).

La descripción básica es una sucesión de ocurrencias del mismo evento en el tiempo, para un mismo individuo (evento recurrente simple) que inicia en $t = 0$, sea $0 \leqslant T_1 < T_2 ...$, los tiempos de ocurrencia del evento, donde T_k es el tiempo de la k-ésima ocurrencia. El proceso de conteo asociado $N(t), 0 \leqslant t$ cuenta el número acumulado de eventos generados durante el proceso. Específicamente, $N(t) = \sum_{k=1}^{\infty} I(T_k \leqslant t)$ es el número de eventos ocurridos en el intervalo $[0, t]$. Más generalmente, $N(s, t) = N(t) - N(s)$ representa el número de eventos ocurridos dentro del intervalo $(s, t]$.

Los modelos para eventos recurrentes son identificados generalmente mediante la consideración de la distribución de probabilidad del número de eventos en intervalos cortos $[t, t+\Delta t]$ dada la historia de la ocurrencia de eventos antes del tiempo t. Para definir cómo se involucra la historia del proceso en el modelamiento, sea $\Delta N(t) = N(t+\Delta t^-) - N(t^-)$ el número de eventos ocurridos en el intervalo $[t, t+\Delta t)$. Dado el supuesto que dos eventos no pueden suceder al mismo tiempo, la *función de intensidad* da la probabilidad instántánea de la ocurrencia de un evento al tiempo t, condicionada a la *historia del proceso* y define el proceso matemáticamente (Cook & Lawless (2006)). Sea $H(t) = N(s) : 0 \leqslant s < t$ la *historia del proceso* hasta el tiempo t, la intensidad (o riesgo) es definida como:

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{P\{\Delta N(t) = 1|H(t)\}}{\Delta t} \tag{1.8}$$

Para el caso multivariado, se considera una serie de modelos indexados por $n = 1, 2, ...$ Además se tiene en cuenta la posibilidad de la observación de tiempos de sobrevida censurados en el seguimiento de los n individuos (en el n-ésimo modelo) de un proceso de conteo multivariado con componente n-ésimo: $N^{(n)} = (N_1^{(n)}, ..., N_n^{(n)})$, donde $N_i^{(n)}$ cuenta los eventos observados en la vida del i-ésimo individuo, $i = 1, ..., n$, sobre el intervalo de tiempo $[0, 1]$(ver Andersen & Gill (1982) para la extensión al intervalo $[0, \infty)$). Así, los caminos muestrales de $N_1^{(n)}, ..., N_n^{(n)}$ son funciones escalonadas, cero en el tiempo cero, con escalones de medida $+1$ solamente. Dos procesos no tienen escalones en el mismo tiempo. Se asume que $N_i^{(n)}(1)$ es casi seguro, finito.

En el modelo multivariado propuesto por Andersen & Gill (1982), las propiedades del proceso estocástico tales como ser una martingala local o un proceso predecible[3] son relativos a la familia de sub-álgebras no decreciente, continuas a la derecha $(H_t^{(n)}: t \in [0, 1])$ en el n-ésimo espacio muestral $(\Omega^{(n)}, H^{(n)}, P^{(n)})$; $H^{(n)}$ representa todo lo que ocurre hasta el tiempo t - la historia del proceso - (en el n-ésimo modelo), $\Omega^{(n)}$ es el espacio de parámetros y $P^{(n)}$ es la medida de probabilidad.

El supuesto básico es que para cada n, $N^{(n)}$ tiene un proceso de intensidad aleatoria $\lambda^{(n)} = (\lambda_1^{(n)}, ..., \lambda_n^{(n)})$, tal que:

$$\lambda_i^{(n)}(t) = Y_i^{(n)}(t)\lambda_0(t) \exp\{\beta_0' Z_i^{(n)}(t)\} \tag{1.9}$$

donde β_0 es un vector columna fijo de p coeficientes, λ_0 es una función de riesgo base y $Y_i^{(n)}$ es un proceso predecible, tomando valores en el intervalo $[0, 1]$ indicando -cuando toma el valor 1- que el i-ésimo individuo está en observación (así, en particular, $N_i^{(n)}$ solo salta cuando $Y_i^{(n)} = 1$).

Finalmente, $Z_i^{(n)} = (Z_{i1}^{(n)}, ..., Z_{ip}^{(n)})'$ es un vector columna de p covariables del proceso para el i-ésimo individuo. Se supone que $Z_i^{(n)}$ es predecible y localmente acotado (Andersen et al. (1993), pág 64).

Al establecer que $N^{(n)}$ tiene proceso de intensidad $\lambda^{(n)}$ se quiere decir que el proceso $M_i^{(n)}$ definido por

$$M_i^{(n)}(t) = N_i^{(n)}(t) - \int_0^t \lambda_i^{(n)}(u)du, \quad i = 1, ..., n, \quad t \in [0, 1] \tag{1.10}$$

son martingalas locales sobre el intervalo de tiempo $[0, 1]$, de hecho son martingalas locales cuadráticamente integrables, de acuerdo con

$$\langle M_i^{(n)}, M_i^{(n)} \rangle(t) = \int_0^t \lambda_i^{(n)}(u)du, \quad y \quad \langle M_i^{(n)}, M_j^{(n)} \rangle = 0, \quad i \neq j, \tag{1.11}$$

[3]Una definición de proceso estocástico predecible y martingala local se encuentra en Andersen et al. (1993), pág 64

donde $\langle M^{(n)}, M^{(n)} \rangle$ es el proceso de variación predecible[4] de M.

Es decir, $M_i^{(n)}$ y $M_j^{(n)}$ son ortogonales cuando $i \neq j$. Sea $\Delta N_i^{(n)}(t) = N_i^{(n)}(t + \Delta t^-) - N_i^{(n)}(t)$ y sobre varias condiciones de regularidad, las cuales no consideramos en profundidad aquí, estos hechos son equivalentes a la siguiente generalización de (1.9):

$$\lambda_i^{(n)}(t|H_t^{(n)}) = \lim_{\Delta t \to 0} \frac{P\{\Delta N_i^n(t) = 1|H_t^{(n)}(t)\}}{\Delta t} \tag{1.12}$$

Lo cual garantiza el modelamiento de tiempos de sobrevida multivariados bajo el modelo de Cox, involucrando los procesos de conteo multivariados, mediante las martingalas[5].

1.2. Modelos para datos de sobrevida multivariados

Los datos de sobrevida multivariados se entienden como múltiples tiempos de sobrevida para un mismo sujeto. En la literatura también se encuentran referenciados como *historia de eventos* o *ciclo de vida* del sujeto (Blossfeld, Golsh & Rohner (2007), Aalen, Borgan & Gjessing (2008)). Dos asuntos importantes se tratan en la incursión de estos métodos: el manejo de la estructura de dependencia de la información ya sea entre los individuos o entre los tiempos y las diversas maneras en las que se pueden presentar los eventos. Un grupo especial de modelos multivariados de sobrevida son los modelos de estado.

1.2.1. Modelos de estado

En un estudio de sobrevida, los eventos a los cuales están expuestos los sujetos se pueden presentar de diferentes maneras. Por ejemplo, un individuo puede pasar por diferentes estados civiles: soltero, casado, separado, viudo; puede estar sano o enfermo, algunos sujetos a lo largo de su historia de vida pueden morir por diferentes causas: por infarto al corazón, diabetes, hipertensión, derrames cerebrales, etc. A los eventos también se les denomina estados. En un estudio de severidad de la enfermedad por ejemplo, un individuo puede pasar por varios estados o bien entrar en algún estado y no volver a ninguno de los demás. Los individuos que sufren de diabetes, en el proceso de cuidado de su enfermedad, por lo general inicialmente, pasan por hacer una dieta estricta, luego deben aplicarse sustitutos del azúcar, en una etapa posterior están expuestos a perder miembros de su cuerpo y posteriormente a caer en coma y finalmente morir. En este proceso se identifican 5 estados; cada uno conlleva al estado siguiente (generalmente estos estados no son reversibles. Es decir, no se vuelve al estado anterior) y finalmente, hay un estado en el que definitivamente no se puede salir que es la muerte. Cuando un estado es tal que los individuos cuando entran a él no pueden pasar a algún otro estado, se le denomina estado absorbente o terminal. A los modelos que se ajustan para analizar este tipo de tiempos hasta la ocurrencia de eventos sobre un mismo individuo, se les denomina modelos de estado o modelos multiestado.

[4] $\langle M^{(n)}, M^{(n)} \rangle(t) = \int_{0<s\leqslant t} E(dM^{(n)}(s)^2|H_s^-)$, (Ver Andersen et al.(1993), Pág 68)

[5] Para una extensión mayor de este resultado, ver Dolivo (1974, Teorema 2.5.1), Aalen (1978, sección 3.2) y Gill (1980, sección 2.3). Las propiedades asintóticas de $\hat{\beta}$ y $\hat{\Lambda}$ se demuestran en Andersen & Gill (1982).

Como los eventos suceden uno luego del otro, (es decir, existe una relación de orden en los tiempos observados hasta la ocurrencia de cada evento para cada individuo) generalmente se presenta el caso de que el último tiempo sea el censurado y esto hace posible el estudio de la secuencia de eventos usando el condicionamiento sucesivo. Para ilustrar, según Rodríguez (2005), considérense tres tiempos consecutivos para la ocurrencia sucesiva de un evento. La distribución conjunta de T_1, T_2 y T_3

$$f_{123}(t_1, t_2, t_3) \tag{1.13}$$

siempre puede ser escrita como el producto de la marginal de T_1, la distribución condicional de T_1 dado T_2, y la distribución condicional de T_3 dado T_1 y T_2:

$$f_{123}(t_1, t_2, t_3) = f_{t_1} f_{2|1}(t_2|t_1) f_{3|12}(t_3|t_1, t_2) \tag{1.14}$$

La contribución a la función de verosimilitud, teniendo en cuenta el hecho que solamente el último tiempo por individuo puede ser el censurado es:

$$f_{t_1} f_{2|1}(t_2|t_1) \lambda_{3|12}(t_3|t_1, t_2)^{d_3} S_{3|12}(t_3|t_1, t_2) \tag{1.15}$$

con d_3 la indicadora que el individuo se encuentra en riesgo y el ultimo término la función de sobrevida condicional para casos censurados y la densidad condicional para el evento muerte.

El modelo de sobrevida de la forma más reconocida, en donde se analiza una variable aleatoria T que representa el tiempo desde un origen determinado hasta la ocurrencia de un evento, es denominado el modelo de dos estados o modelo de mortalidad donde el sujeto inicialmente está vivo (estado 0) y luego pasa a estar muerto (estado 1)(figura 1.1). Según Andersen & Keiding (2002) , se puede observar que para la variable aleatoria T, $S(t)$ y $F(t)$, corresponden a las probabilidades estar en el estado 0 o 1 en el tiempo t, respectivamente. Si se supone que cada individuo está en el estado 0 en el tiempo 0 entonces $F(t)$ es también la probabilidad de transición del estado 0 al estado 1 en el intervalo de tiempo de 0 a t. En tiempo continuo, la distribución de T además de ser caracterizada por $S(t)$ y $F(t)$, lo es también por la función de tasa de riesgo:

$$\alpha(t) = -d \log S(t)/dt = \lim_{\delta t \to 0} \frac{P(T \leqslant t + \delta t | T \geqslant t)}{\delta t} \tag{1.16}$$

es decir que,

$$S(t) = \exp(-\int_0^t \alpha(u) du) \tag{1.17}$$

Así, $\alpha(.)$ es la intensidad de transición del estado 0 al estado 1, corresponde a la probabilidad instantánea por unidad de tiempo de ir del estado 0 al estado 1.

En general, los análisis de los tiempos en que suceden eventos de interés sobre los individuos, se centran en hacer inferencias para la intensidades y probabilidades de transición en los modelos multiestado. Esto incluye la estimación de los coeficientes de los modelos propuestos y las respectivas pruebas de hipótesis para estos mismos estimadores. Tener en

cuenta la posible dependencia de los tiempos y covariables asociadas a la ocurrencia de los eventos, resulta ser entre otros, los puntos de mayor interés en los análisis de sobrevida asociados a este tipo de información.

Como en el análisis de sobrevida univariado, en los modelos multiestado la censura se presenta tanto a izquierda como a derecha. Por ejemplo, no todos los individuos en observación llegan a un estado absorbente, aquellos que no llegaron resultan ser censurados a derecha. Ahora, cuando el inicio de la observación no es el mismo para todos los individuos, sino que, puede haberse dado que algunos de ellos hayan ingresado después y hayan experimientado el evento de interés antes de ingresar al proceso de observación de los eventos, tales individuos son censurados a izquierda.

Finalmente también, y no menos importante, se le debe poner especial atención a garantizar en este tipo de modelos, que sea factible la censura de los individuos de manera independiente de la observación de los tiempos en el proceso,(se denomina censura independiente) pues esto garantiza que el análisis sea *representativo* para la población sin censura. Esto significa que los individuos que son censurados, no deberían tener un riesgo más alto ni más bajo de eventos futuros que los que los otros.

A continuación se presentan la definición y los diferentes tipos de modelos de estado:

Un proceso multiestado, es un proceso estocástico $(X(t), t \in \Im)$ con un *espacio estado* finito $\boldsymbol{S} = 1, ..., p$ y con trayectorias de la muesta continuas a derecha: $X(t+) = X(t)$. Aquí, $\Im = [0, \tau]$ o $(0, \tau]$ con $\tau \leqslant +\infty$. El proceso tiene una distribución inicial $\varpi_h(0) = P(X(0) = h,\ h \in \boldsymbol{S}$. Un proceso multi-estado $X(\cdot)$ genera una historia $\boldsymbol{H}_t$ (una σ - álgebra) que consiste en las observaciones del proceso en el intervalo $[0, t]$. Respecto a esta historia se definen las probabilidades de transición como:

$$P_{hj}(s,t) = P(X(t) = j | X(s) = h, \boldsymbol{H}_{s^-}) \tag{1.18}$$

para h,$j \in \boldsymbol{S}$, $s, t \in \Im$, $s \leqslant t$ y se definen las intensidades de transición mediante:

$$\alpha_{hj} = \lim_{\delta t \to 0} \frac{P_{hj}(t, t+\delta t)}{\delta t} \tag{1.19}$$

las cuales se asume que existen. Algunas intensidades de transición pueden ser cero para todo t. Como se observa en la figura 1.1, los modelos de estado pueden ser presentados usando diagramas con cajas representando los estados y con flechas entre los estados representando las posibles transiciones, es decir, las intensidades de transición diferentes a cero. Un estado $h \in \boldsymbol{S}$ es absorbente si para todo $t \in \Im$, $j \in \boldsymbol{S}$, $j \neq h$, $\alpha_{hj} = 0$; de otra manera, h es transitorio. Las probabilidades de estado $\varpi_h(t) = P(X(t) = h)$ están dadas por:

$$\varpi_h(t) = \sum_{j \in \boldsymbol{S}} \varpi^j(0) P_{jh}(0,t) \tag{1.20}$$

Obsérvese que la $P_{jh}(\cdot, \cdot)$ y así también $\alpha_{hj}(0)$ dependen tanto de la medida de probabilidad

como de la historia. Esta dependencia ha sido suprimida de la notación. Si $\alpha_{hj}(0)$ solo depende en la historia del estado $h = X(t)$ en t, entonces el proceso es Markoviano[6].

1.2.1.1. Modelo para datos de sobrevida de dos estados

El modelo de dos estados como su nombre lo indica solo tiene en cuenta dos estados: vivo-muerto, empleado-desempleado, con hijos-sin hijos, etc. El más conocido es el ilustrado en la figura 1.1. Tal modelo es denominado modelo de mortalidad. Tiene $p = 2$ estados y solo una posible transición del estado 0 al estado 1. La intensidad de transición correspondiente $\alpha_{0,1}(t)$, está dada por la función de tasa de riesgo $\alpha(t)$, mientras que $\alpha_{1,0}(t) = 0$ para todo t, esto es, el estado 1 es absorbente. La distribución inicial es degenerada [7] en 0: $\varpi_0(0) = 1$ y el proceso es Markoviano .

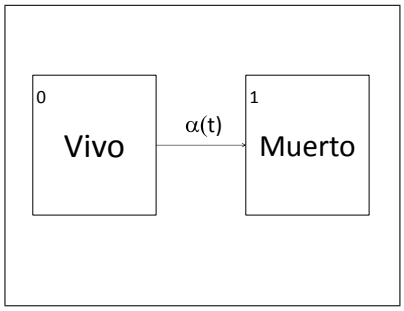

FIGURA 1.1. Modelo mortalidad o de dos estados.

1.2.1.2. El modelo de riesgos en competencia

Este modelo tiene un estado transitorio "0 : *vivo*"y un número k de estados absorbentes, el estado h, $h = 1, ..., k$ corresponde a la "muerte por la causa h". Así hay $p = k + 1$ estados (figura 1.2).

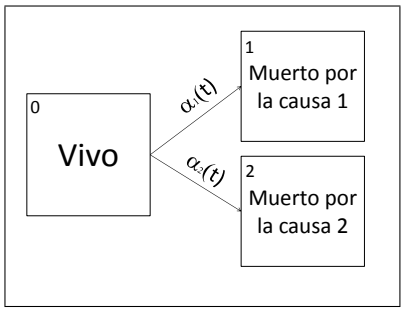

FIGURA 1.2. Riesgos en competencia para mortalidad por dos causas.

[6]un proceso estocástico de $E_1(t), E_2(t), ..., E_K(t)$, estados, ($E_i(t)$ representa la ocurrencia del estado i en el tiempo t. $t_1 < t_2 < t_n$) se dice Markoviano cuando $P(E_K(t)|E_1(t), E_2(t), ..., E_{K-1}(t)) = PE_K(t)|E_{K-1}(t)$

[7]Una v.a X se dice que tiene una distribución degenerada en un punto s si su función de masa es $P_X(x) = \{1 \; si \; x = p; \; 0 \; si \; x \neq p\}$

Las intensidades de transición $\alpha_{0,h}(t)$ para $h = 1, ...; k$ están dadas por las funciones de riesgo de "causa específica ":

$$\alpha_h(t) = \lim_{\delta t \to 0} \frac{P(Muerto\, por\, la\, causa\, h\, en\, t + \delta t | T \geqslant t)}{\delta t} \tag{1.21}$$

Donde T es el tiempo de sobrevivencia. La distribución inicial es degenerada en el estado 0 (para el caso de la figura 1.2, el estado *vivo*), el único estado transitorio del modelo, es decir, $\alpha_{hj}(t) = 0$ para todo $h \neq 0$ y todo j. Las probabilidades de transición están dadas por la función de sobrevida:

$$P_{00}(0,t) = S(t) = P(T > t) = \exp(-\int_0^t \sum_{h=1}^{k} \alpha_h(u)du) \tag{1.22}$$

y las funciones de incidencia acumuladas:

$$P_{0h}(0,t) = \int_0^t S(u-)\alpha_h(u)du, h = 1, ..., k \tag{1.23}$$

Como el modelo de dos estados (k=1) el modelo de riesgos en competencia es Markoviano.

1.2.1.3. El modelo de enfermedad - muerte

En este modelo frecuentemente el tiempo t es la edad del individuo y usualmente se asume que los individuos están en el estado 0 en $t = 0$ (figura 1.3). Sin embargo los individuos no siempre son observados desde $t = 0$. La mortalidad α_{12} del enfermo, algunas veces puede depender de la duración d desde la entrada al estado 1 y adicionalmente de la dependencia de la "edad"t. Si α_{12} no depende de d, el proceso es Markoviano, de otra manera es un proceso semi-Markoviano [8].

En la figura 1.3 se tiene en cuenta la posibilidad de reversibilidad: es decir, la transición de vuelta del estado 1 al estado 0 es posible. Cuando no se tiene en cuenta esta posibilidad el modelo se denomina unidireccional. Un ejemplo en el cual se puede pensar en un modelo con reversibilidad es el caso de cáncer, cuando la terapia a la que se somete el individuo puede llegar a eliminar por completo la enfermedad. Un ejemplo de modelo unidireccional es el de enfermedades definidas como terminales (Alzheimer, Parkinson, etc).

Así,las probabilidades de transición en este modelo tienen las siguientes expresiones:

$$P_{00}(s,t) = \exp(-\int_s^t (\alpha_{02}(u) + \alpha_{01}(u))du) \tag{1.24}$$

y (en el caso Markoviano)

[8] Un proceso semi-markoviano es un proceso estocástico en tiempo continuo $\{E(t), t \geq 0\}$; donde en cada transición a un estado i en un tiempo s, se cumple que $E(s + t)$ es independiente de $E(u)$, $u < s$, para todo t, u y para cualquier transición

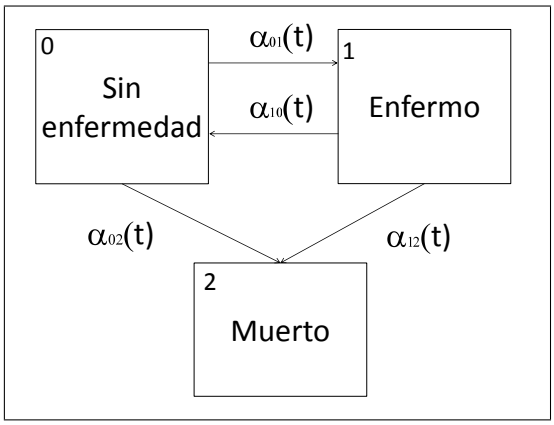

FIGURA 1.3. Modelo gráfico de enfermedad- muerte.

$$P_{01}(s,t) = (\int_s^t P_{00}(s,u-)\alpha_{01}(u)P_{11}(u,t)du) \tag{1.25}$$

de donde

$$P_{11}(s,t) = \exp(-\int_s^t \alpha_{12}(u)du) \tag{1.26}$$

1.2.1.4. El modelo de eventos recurrentes

Si el interés es analizar la ocurrencia reiterativa de un evento dado sobre una misma unidad o individuo (figura 1.4), por ejemplo la reincidencia de un tumor cancerígeno, el número de ataques epilépticos en un paciente, número de embarazos de una mujer, etc; un modelo como el que se describe en la figura 1.4 puede ser el adecuado (en la figura no se grafica un evento absorbente o terminal , que puede ocurrir). En este tipo de datos es frecuente el interés de estimar el número esperado de eventos ocurridos en el intervalo $[0,t]$. En la siguiente sección se hace una presentación más amplia de este tipo de modelos.

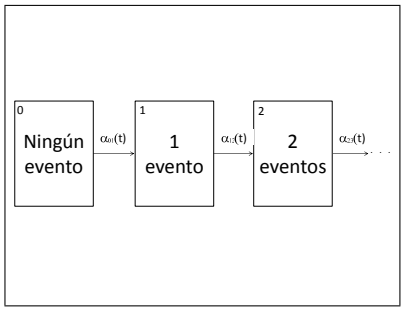

FIGURA 1.4. Modelo gráfico para eventos recurrentes

Un apunte final a esta sección es que la definición de qué tipo de modelo de estado usar es tan importante como definir si el modelo a ajustar es lineal o cuadrático. La idea

última de los modelos de estado es no ignorar la presencia de un proceso subyacente que tiene relación con el evento de estudio.

1.2.2. Eventos recurrentes

Un evento recurrente es aquel que sucede en varias ocasiones para un mismo individuo. En ciencias de la salud, se reconocen como eventos recurrentes por ejemplo los episodios de asma, los infartos no letales al corazón, la incidencia del cáncer luego del tratamiento para eliminarlo, crisis epilépticas. En la industria, se identifica la recurrencia de falla de una máquina luego de ser reparada, en la economía un evento recurrente es la caída del dólar, la recesión, entre otros.

El modelamiento de eventos recurrentes, implica la consideración de los diferentes eventos ocurridos sobre un individuo. Como se presentó en la secciòn 1.1.1.2, una de las herramientas que facilita en gran medida la posibilidad de detallar este proceso, son los procesos de conteo. Varios autores exponen de manera muy explícita y detallada la manera como se entienden los modelos de sobrevida desde el punto de vista de los procesos de conteo (Therneau & Grambsch (2000), Andersen et al. (1993)).

Basados en el contexto teórico expuesto en la sección 1.1.2, en esta sección se presenta una parte de la fundamentación matemática para el análisis de tiempos de eventos recurrentes.

Sea $N(s,t)$ el número de ocurrencias de algún tipo de evento sobre el intervalo de tiempo $(s,t]$ para un individuo. Por conveniencia se asume que el proceso inicia en $t=0$ con $N(0)=0$ y se define $N(t)=N(0,t)$ para $t>0$. El proceso $\{N(t), 0\leq t\}$ es entonces el proceso de conteo para la recurrencia de los eventos. Para este trabajo se considera un solo evento de interés que es la desinstalación de alguno de los servicios de telecomunicaciones.

Dado que el análisis se centra en tiempos hasta un evento y para el caso de las recurencias los eventos suceden en diferentes momentos del tiempo, es necesario definir la mecánica de registro de los tiempos a analizar. Para el caso de este trabajo se usa el tiempo que sucede desde el inicio del seguimiento y hasta el primer evento, luego se renueva el tiempo tomando para la segunda recurrencia el tiempo entre la primera y la segunda. Este tipo de tiempos se conocen como tiempos de espera o "*gap times*". Se denota $B_j = T_j - T_{j-1}$, como el tiempo de espera entre el $(j-1)$ y el j-ésimo evento, o tiempo inter-ocurrencias.

Se supone que en el caso de tiempo contínuo dos eventos no pueden ocurrir simultáneamente y la función de riesgo (función de intensidad) para el proceso del evento está definida como en 1.28:

$$\lambda(t|H(t)) = \lim_{\Delta t\to 0} \frac{P\{\Delta N(t)=1|H(t)\}}{\Delta t} \tag{1.27}$$

Se asume que la función de intensidad es acotada y continua, excepto posiblemente en un número finito de puntos fuera (a la derecha) de cualquier intervalo de tiempo finito. La función de intensidad define un proceso de evento, y todas las características del proceso pueden ser determinadas a partir de ella

1.2.3. Modelos para eventos recurrentes

En las propuestas para modelar eventos recurrentes el modelo de Cox toma su relevancia. Un asunto importante a considerar en la extensión del modelo de riesgos proporcionales de Cox al modelamiento de tiempos a eventos recurrentes es la correlación intra-sujeto. Para un manejo adecuado de este supuesto se conocen los modelos marginales. En esta sección se hace una breve descripción de tales modelos, según la presentación de Therneau & Grambsch (2000) .

Varianza Robusta

Cuando un sujeto contribuye a la estructura de datos con varios eventos, el supuesto de independencia de las observaciones del modelo de Cox no se tiene. El modelo de varianza robusta, en donde Lipsitz et al. (1990) propone un mecanismo de estimación de la varianza de $\hat{\beta}$ mediante una corrección basada en un estimador Jacknife agrupado, se propone como alternativa para considerar el supuesto del modelo de Cox de independencia de las observaciones. Los valores del estimador Jacknife agrupado se definen como $J_i = \hat{\beta} - \hat{\beta}_{(i)}$, donde $\hat{\beta}_{(i)}$ es resultado del ajuste que incluye a todos los individuos excepto al individuo i. Es denominado agrupado porque en el caso de múltiples eventos, un individuo aporta varias observaciones y eliminar un indiviuo implica eliminar un grupo determinado de observaciones. Therneau & Grambsch (2000) describen una forma de calcular los valores del Jacknife agrupado directamente en la iteración de Newton-Raphson. El cambio en el vector de coeficientes estimado se puede encontrar haciendo $\Delta\beta = 1'(U\mathcal{I}^{-1}) \equiv 1'D$, donde D es la matriz de residuos. Entonces el cambio en $\hat{\beta}$ en cada iteración es la suma de columnas de la matriz D definida como la puntuación residual escalada por $\mathcal{I}^{-1}$ (la varianza de $\hat{\beta}$).

Este estimador Jacknife agrupado puede ser usado para obtener estimaciones robustas de la varianza para el modelo de Cox. Si J es la matriz de valores agrupados Jacknife (es decir, la i-ésima fila de J es $\hat{\beta}-\hat{\beta}_{(i)}$), entonces el estimador Jacknife agrupado de la varianza se puede escribir como el producto $V_j = \frac{n-1}{n}(J-\bar{J})'(J-\bar{J})$, donde $\bar{J}$ es la matriz columna de medias de J. Una aproximación natural es $D'D$, la matriz producto de las variazas aproximadas Jacknife (ignorando el término $\frac{n-1}{n}$). Escribiendo $D'D = \mathcal{I}^{-1}(U'U)\mathcal{I}^{-1}$, esta varianza puede ser vista como un estimador emparedado ABA donde A es la varianza usual y B es un término de corrección. Además de insesgado, este estimador Jacknife agrupado es comúnmente más variable que la varianza típica del modelo de Cox pero es una varianza robusta que trata adecuadamente la correlación dada por los eventos repetidos por individuo y por lo tanto se espera que informe cuando se ajusten los modelos marginales.

El modelo de eventos ordenados, que supone que los eventos/estados se presenten con un orden lógico para el fenómeno en seguimiento; por ejemplo en Castañeda & Gerritse (2010) se presenta el análisis de un caso de hospitalizaciones recurrentes, es un ejemplo típico de eventos ordenados pues se definen los eventos para que se presenten de manera secuencial: luego de la hospitalización 1 viene la hospitalización 2 y luego las demás hospitalizaciones hasta la muerte. Ahora, acercamientos más comunes a este tipo de modelos de eventos ordenados, son los de incrementos independientes (Andersen & Gill, 1982) , marginal (Wei et al. 1989) , y PWP (Prentice et al. 1981) . Todos los mencionados enmarcados dentro del contexto teórico de modelos de regresión marginal en donde $\hat{\beta}$ se

determina mediante un ajuste que no tiene en cuenta la correlación entre los eventos, seguido de una corrección de la varianza y difieren considerablemente en la construcción de los conjuntos de riesgo.

Modelo de Andersen y Gill (AG)

Este modelo entre todos los marginales, es el más simple, es muy cercano a un modelo de regresión de Poisson. En este modelo el proceso de intensidad para el sujeto i es:

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{Z_i(t)\beta\} \tag{1.28}$$

La diferencia con el modelo de Cox radica en la variable indicadora de riesgo $Y_i(t)$. En el modelo de Cox, la i-ésima unidad termina de estar en riesgo una vez que el evento le haya ocurrido por primera vez al individuo y por tanto $Y_i(t)$ pasa de tomar el valor 1 a tomar el valor 0, mientras que en el modelo de AG (Andersen & Gill, (1982)) para eventos recurrentes $Y_i(t)$ permanece igual a 1 cuando el evento ocurre sobre la misma unidad i. El modelo de AG es apropiado para situaciones en las que los eventos observados sobre una misma unidad se pueden asumir mutuamente independientes.

Modelo de Wei, Lin y Weissfeld(WLW)

Wei, Lui y Weissfel (1989) proponen métodos semiparamétricos para analizar tiempos de falla multivariados. En este modelo la salida ordenada de tiempos de los eventos es tratada como si fuera un caso de riesgos en competencia. En el análisis se conforman estratos por el orden de ocurrencia del evento y el número de estratos es igual al número máximo de eventos reportado por los individuos en el estudio.

Sea T_{ij} el tiempo para la j-ésima ocurrencia del evento sobre la i-ésima unidad,con $j = 1, ..., K$ y $i = 1, ..., n$; K es el máximo número de eventos observado en los datos. Sea $Z_{ij} = (Z_{ij,1}(t), ..., Z_{ij,p}(t))^{'}$ un vector de covariables para la i-ésima unidad en el tiempo $t \geqslant 0$ respecto al j-ésimo evento. En el modelo AG no se considera dentro del modelamiento la estructura de dependencia intra-sujeto debida a la recurrencia de los eventos por sujeto. Los modelos de WLW son propuestos para modelar funciones de riesgo marginal mediante funciones de intensidad condicionadas a un proceso de conteo $N_i(t)$. Para la j-ésima ocurrencia del evento sobre la i-ésima unidad, se asume que la función de riesgo $\lambda_{ij}(t)$ toma la forma:

$$\lambda_{ij}(t) = Y_{ij}(t)\lambda_{0j}(t)\exp\{Z_{ij}^{'}(t)\beta_j\} \tag{1.29}$$

Como K es el número máximo de eventos sobre alguna de las unidades, es natural que si una unidad le ocurre un número de eventos L menor que K éstas tendrán valores faltantes sobre la ocurrencia del evento después de la L-ésima ocurrencia. Este modelo permite desarrollar un análisis por separado para cada estrato j y para cada interacción estrato y covariable. El indicador de riesgo, Y_{ij} , es igual a 1 hasta la ocurrencia del j-ésimo evento, a menos que la unidad sea censurada. Ninguna estructura particular de dependencia entre los tiempos de falla en cada unidad es impuesta. Los parámetros de

regresión son estimados mediante la respectiva función de verosimilitud parcial.

Modelo Prentice, William y Peterson (PWP)

El modelo de Prentice, William y Peterson (1981), define claramente el orden de ocurrencia de los eventos. Un sujeto no se encuentra en riesgo para el k-ésimo evento si no ha experimentado el evento anterior (k-1). Ellos consideran dos clases generales de modelos de regresión para eventos recurrentes los cuales relacionan el riesgo como una función de intensidad con covariables y la historia de falla. Los modelos consideran e incluyen el tiempo desde el origen del estudio hasta la ocurrencia de cada evento y el tiempo interocurrencia, respectivamente. Ambos modelos son estratificados de riesgos proporcionales, lo que significa que la función de intensidad puede variar de un evento a otro, mientras que en el modelo de AG se asume que todos los eventos son idénticos.

Sean $Z(u) = Z_1(u), ..., Z_p(u)$ un vector de covariables, para un sujeto bajo estudio, el cual está bajo observación en el tiempo $u \geqslant 0$, y $Z(t) = \{z(u) : u \leqslant t\}$ el correspondiente proceso de covariables hasta el tiempo t. Similarmente, sea $N(t) = \{N(u) : u \leqslant t\}$, donde $N(u)$ es el número de eventos ocurridos antes del tiempo u. La función de riesgo o de intensidad al momento t -que es definida como una tasa instantánea de riesgo al momento t- dadas las covariables y el proceso de conteo en el tiempo t es

$$\lambda\{t|N(t), Z(t)\} = \lim_{\delta t \to 0} \frac{Pr\{t \leqslant T_{n(t)+1} < t + \delta t|N(t), Z(t)\}}{\delta t} \tag{1.30}$$

formulación que, siguiendo a Cox (1972) y a Therneau & Grambsch (2000), se puede escribir como el producto entre una función arbitraria y una función exponencial en las covariables. Ellos presentan dos tipos de funciones de línea base, una en función del tiempo desde el inicio del estudio hasta el momento t, y otra desde el evento inmediatamente anterior, $t - t_{n(t)}$. Además, parece conveniente permitir que la forma de la función de riesgo dependa del número de eventos anteriores y posiblemente de otras características, esto se condensa en $\{N(t), Z(t)\}$. Así, se dispone de dos modelos de riesgo parcialmente paramétricos:

$$\lambda\{t|N(t), Z(t)\} = \lambda_{0s}(t) \exp\{Z_t^{'}(t)\beta_s\} \tag{1.31}$$

y

$$\lambda\{t|N(t), Z(t)\} = \lambda_{0s}(t - tn(t)) \exp\{Z_t^{'}(t)\beta_s\} \tag{1.32}$$

donde para ambos casos $\lambda_{0s}(.) \geqslant 0$ $s = (1, 2, ...)$ son funciones de riesgo base arbitrarias, la variable de estratificación $s = \{N(t), Z(t), t\}$ puede variar como una función de tiempo para un sujeto dado y β_s es un vector columna de los coeficientes de regresión estratificados.

Modelo Chang y Wang (CW)

Chang y Wang (1999) proponen un modelo de riesgo semiparamétrico para tiempos de eventos recurrentes a través de un modelo de regresión condicional utilizando los modelos de Cox y PWP:

$$\lambda_{ij}(t|H(t), Z(t)) = \lambda_{j0}(t - t_{j-1}) \exp\{\beta^{'} Z_{i1}(t) + \gamma_j Z_{i2}(t)\} \tag{1.33}$$

con $H(t)$ la historia del evento hasta el tiempo t. β es el parámetro estructural y γ es un parámetro que está asociado a un evento específico, tales parámetros corresponden a los efectos asociados con los vectores de covaraibles Z_{i1} y Z_{i2} respectivamente. Este modelo es útil cuando el interés se focaliza sobre ocurrencias específicas del evento.

1.2.4. Modelo para eventos recurrentes con evento terminal

La recurrencia de un evento determinado sobre el mismo sujeto puede conducir a un evento final. En medicina, por ejemplo, en casos de cáncer la incidencia de tumores luego de la extirpación del primero o ataques consecutivos de hidrocefalia, normalmente se denotan como recurrencias de un evento en el seguimiento de la enfermedad. Tales recurrencias pueden conducir al evento final que en estos casos es la muerte. Para el caso de los clientes corporativos[9] en empresas de telecomunicaciones, la desconexión paulatina de los servicios contratados, provoca un evento final que es la pérdida definitiva del cliente.

Como se enunció en la sección 1.2.1, en los modelos de estado normalmente se puede diferenciar entre los diferentes estados uno que es el terminal o absorbente. Un estado terminal determina que el sujeto cuando llega a él no vuelve a sufrir ningún otro estado (no sale de allí). Es decir para este individuo termina el proceso, en los dos casos citados de cáncer e hidrocefálea, el evento terminal es la muerte y en el caso de los clientes empresariales en empresas de telecomunicaciones, el evento terminal es la pérdida definitiva del cliente.

Un diagrama simple que representa este proceso es el que se muestra en la figura 1.5.

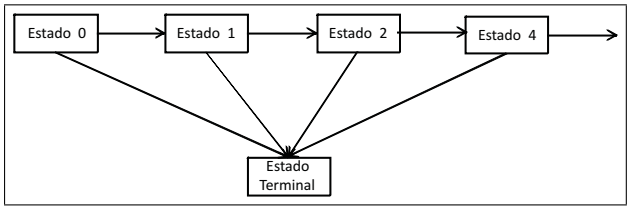

FIGURA 1.5. Eventos recurrentes con un estado terminal.

Para modelar los tiempos de sobrevida considerando eventos recurrentes y un evento terminal, se propone el modelamiento teniendo encuenta modelos separados para los dos procesos: el de recurrencias y el terminal (Cook(2006)). Sea T_i el tiempo del evento terminal para el sujeto i y defínase $D_i(t) = I(t \leqslant T_i)$ y $Y_i(t) = D_i(t)I(t \leqslant C_i)$, donde C_i es la censura correspondiente al final del seguimiento. Si $H_i(t) = \{(N_i(s), D_i(s)) : 0 \leqslant s < t\}$,

[9] clientes grandes del segmento de empresas

($N_i(s)$ es el número de eventos ocurridos hasta el tiempo s) representa la historia del proceso hasta el tiempo t, un modelo completo para el proceso se puede expresar en términos de las funciones de intensidad de eventos: el terminal y los recurrentes, que se modelan como dos tipos de eventos diferentes.

La función de intensidad (riesgo) de los eventos recurrrentes es:

$$\lambda_i(t|H_i(t)) = \lim_{\Delta t \to 0} \frac{Pr\{\Delta N_i(t) = 1|H_i(t)\}}{\Delta t} \tag{1.34}$$

La función de intensidad (riesgo) del evento terminal es:

$$\gamma_i(t|H_i(t)) = \lim_{\Delta t \to 0} \frac{Pr\{T_i < t + \Delta t|H_i(t), D_i(t) = 1\}}{\Delta t} \tag{1.35}$$

Como lo muestra la figura 1.5, el evento terminal se puede interpretar de alguna manera como un proceso de eventos recurrentes "cuasi en competencia"ya que además de que el sujeto se encuentra en riesgo por cualquiera de los eventos recurrentes que experimenta, también lo está por el evento terminal. Si n_i eventos recurrentes son observados en los tiempos $t_{i1}, ..., t_{in_i}$ sobre el i-ésimo individuo en el intervalo $[0, \tau_i]$, donde $\tau_i = min(T_i, C_i)$ y $\delta_i = I(T_i \leqslant C_i)$, entonces sobre censura independiente (la censura no depende de la ocurrencia de los eventos en estudio), la función de verosimilitud es proporcional a:

$$\prod_{j=1}^{n_i} \lambda_i(t_{ij}|H_i(t_{ij}))[\gamma_i(\tau_i|H_i(\tau_i))]^{\delta_i} \times \exp\left(-\int_0^{\tau_i} [\lambda_i(u|H_i(u)) + \gamma_i(u|H_i(u))]du\right) \tag{1.36}$$

Las inferencias pueden basarse en las verosimilitudes parciales derivadas de la factorización de ésta en dos partes.

Ahora, como se ha comentado, el proceso de recurrencia sobre cada uno de los diferentes tipos de eventos puede desembocar en que se finaliza tal recurrencia con el evento terminal. Esto denota una dependencia entre el proceso de las recurrencias sobre los diferentes tipos de eventos y el proceso del evento terminal. Para hacer un modelamiento más adecuado según esta consideración, se incluye en el modelo completo (el modelo conjunto que considera el modelo de riesgo de recurrencias por cada tipo de evento y el del evento terminal) efectos aleatorios que permitan interpretar este nivel de dependencia entre los dos procesos.

1.3. Tratamiento de la censura

La situación de los datos en modelos de eventos recurrentes es que se observa un conjunto de n procesos sobre períodos de tiempo. Normalmente, cada proceso comienza en un tiempo 0 en el estado 1 y se observa hasta algún tiempo $\mathcal{T}$ que es diferente entre procesos. En el caso de procesos con un evento absorbente o terminal no hay información de los demás procesos cuando el sujeto entra en el estado terminal. Cuando el estado terminal no se alcanza, el final de la observación es un tiempo censurado. Para cada proceso son observados un número de eventos digamos E. Los procesos observados pueden ser recolectados como los tiempos de transición $T_1, ..., T_E$ y los estados ingresados serían:

$S_1, ...S_E$. Para describir el período desde T_E hasta $\mathcal{T}$ se introduce otro tiempo, cuando $T_E < \mathcal{T}$, llámese $T_{E+1} = \mathcal{T}$, con variable de estado $S_{E+1} = 0$. Entonces, si se presenta un evento en el tiempo $\mathcal{T}$, el número de tiempos K es igual a E, y cuando no se presenta el evento, el último tiempo es censurado y $K = E+1$. Los tiempos de transición se notan por $T_1, ..., T_K$ y los estados de tales transiciones son $S_1, ..., S_K$. En el caso de censura $S_K = 0$. S_0 es el estado inicial.

1.4. Modelos de fragilidad

Los modelos de fragilidad han dirigido su atención hacia el análisis de la información acerca de la historia de ocurrencia de un evento a través del modelo de Cox y varias de sus extensiones. Un modelo de fragilidad es un modelo de riesgo multiplicativo que consta de tres factores: un término de fragilidad (efecto aleatorio), una función de riesgo base (paramétrica o no paramétrica) y un término que considera la influencia de algunas covariables observadas (efectos fijos).

El aporte de esta clase de modelos es que consideran dos hechos importantes en el análisis de la historia de ocurrencia de los eventos: *i*) La situación de una heterogeneidad presente entre los individuos y *ii*) La consideración de un proceso de dependencia subyacente entre los tiempos de ocurrencia del mismo evento. Dado que en los modelos univariados todas las duraciones describen el tiempo al mismo tipo de eventos, los tiempos de los eventos son considerados como independientes.

1.4.1. Modelos de fragilidad univariados

Evaluar la heterogeneidad e incluirla en el modelo no es fácil pero sí muy importante. Esta situación es la que normalmente se trata en los modelos de fragilidad univariados. La principal idea de los modelos de fragilidad es proporcionar una manera de introducir efectos aleatorios que permitan entender la variabilidad no observable en los datos de sobrevida. La idea de captar y poder entender cómo varía de individuo a individuo su curva de riesgo, es ambiciosa y no es de desconocimiento de ningun investigador que poder incluir los factores asociados al riesgo, ya sea mediante el efecto de las covariables o algún otro factor puede ser tan dificil como costoso. Normalmente en los datos de sobrevida se tiene muy poca información y llegar a ella frecuentemente no es viable. En tal caso es útil tener en cuenta dos fuentes de variabilidad en los datos de tiempo hasta un evento: variabilidad tenida en cuenta por factores de riesgo observables incluidos en el modelo (y teoricamente predecibles) y la heterogeneidad causada por covariables desconocidas, las cuales son teóricamente impredecibles. Esta forma de tener en cuenta la variabilidad en este tipo de datos la cubren los modelos mixtos mediante la inclusión de una variable en donde se puede entender el efecto del impacto de las covariables a nivel de individuo. Así, los riesgos no observados son descritos por el coeficiente de tal variable, la cual es llamada en análisis de sobrevida, ***fragilidad*** . Esta es una variable aleatoria que se asume sigue alguna distribución.

Es posible hacer diferentes escogencias de la distribución para las covariables no observadas. La popularidad del modelo de fragilidad log-normal se deriva principalmente de la relación con los modelos mixtos generalizados, donde el supuesto habitual es que los efectos aleatorios siguen una distribución normal. Sobre la escogencia de la distribución

del término de fragilidad se determina el grado de heterogeneidad no observada, lo que no es posible conseguir en un modelo de riesgos proporcionales.

En teoría, si hay fragilidades no medidas o no observadas, la razón de riesgo, no solamente será una función de las covariables, sino que también debe ser función de las fragilidades. De esta manera el modelo (1.3) se puede escribir como:

$$\lambda_i(t) = \lambda_0(t) exp(\beta^{'} Z_i + \psi^{'} w_i) \tag{1.37}$$

donde w_i son las fragilidades que se asumen independientes provenientes de una distribución con media 0 y varianza 1 (Klein & Moeschberger (1997)). Se observa que este modelo tiene la forma clásica del modelo mixto, donde se consideran efectos aleatorios (en w_i) y efectos fijos (en β).

Dos aspectos importantes de este modelo para tener en cuenta son:

- Si $\psi = 0$ entonces se tiene el modelo de riesgos proporcionales usual.
- Si los valores relevantes de w_i estuvieran medidos dentro del modelo (o se pudieran medir), entonces ψ debería tender a 0

Así, de (1.38) se puede derivar un modelo para tener en cuenta la heterogeneidad no observada. Se deben algunos supuestos sobre la distribución del término de fragilidad. Para probar esto, se reescribe (1.38) de la siguiente manera:

$$\lambda_i(t|\beta^{'} Z_i, \nu_i) = \lambda_0(t) \nu_i \exp(\beta^{'} Z_i) \tag{1.38}$$

nótese que $\nu_i = \exp(\psi^{'} w_i)$.

Mediante esta formulación, se da cuenta de cómo las fragilidades actúan multiplicativamente sobre el riesgo. Para propósitos de identificabilidad, se asume que la media de ν es 0 y la varianza es desconocida e igual a algún parámetro θ.

Si el riesgo es una función de fragilidades, la función de sobrevida debe estar también condicionada sobre ambos: las covariables y el término de fragilidad. La función de sobrevida condicionada (omitiendo subíndices) está dada por:

$$\begin{aligned} S(t|\beta^{'} Z, \nu) &= \exp(-\int_0^t \lambda(u|\nu) du) \\ &= exp(-\nu \int_0^t \lambda(u) du) \end{aligned}$$

1.4.2. Modelos de fragilidad multivariados

Para presentar una discusión del segundo tema de interés enunciado para el análisis de tiempos hasta un evento, hay que decir que los modelos de fragilidad (mediante el modelamiento de un término de fragilidad) también pueden ser usados para modelar asociaciones entre los tiempos hasta un evento. Esta característica es usada cuando se tienen

varios tiempos para un mismo individuo (caso multivariado), en donde los tiempos pueden estar asociados. Un caso para mencionar, es el seguimiento de eventos recurrentes donde la ocurrencia del primer evento puede tener impacto sobre el segundo y así sucesivamente, lo que implica un nivel de asociación de estos tiempos observados en cada individuo. Usualmente los individuos se toman como conglomerados [10] y el modelo se denomina modelo de "fragilidad compartida".

El modelo en este caso tiene en cuenta los "conglomerados"conformados por las j observaciones cada uno y se escribe como:

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\beta^{'} Z_{ij} + \psi^{'} w_i) \tag{1.39}$$

donde los w_i son fragilidades de los clusters o subgrupos, las cuales se asumen independientes y provenientes de una distribución con media 0 y varianza θ.

También como se dispuso para los modelos de fragilidad univariados, este modelo se puede expresar de la siguiente manera:

$$\lambda_{ij}(t|\beta^{'} Z_i j, \nu_i) = \lambda_0(t)\nu_i \exp(\beta^{'} Z_i j) \tag{1.40}$$

donde $\nu_i = \exp(\psi^{'} w_i)$, éstas son fragilidades compartidas para los individuos, es decir, para los conglomerados.

Nótese la diferencia entre esta expresión y la presentada en (1.39). Aquí la fragilidad es compartida entre las j observaciones para cada conglomerado/subgrupo que para eventos recurrentes generalmente son las unidades/sujetos/individuos sobre las que se tienen los tiempos de las recurrencias de los eventos en estudio.

Este es un modelo mixto porque el riesgo común en cada "conglomerado"se supone que es aleatorio. El modelo asume que todos los tiempos en un "conglomerado"son independientes dadas las variables de fragilidad. En otras palabras, este es un modelo de independencia condicional donde la fragilidad es común a todos los individuos en un "conglomerado"y por lo tanto dan cuenta de la dependencia entre los tiempos hasta el evento. Esta es la razón del concepto fragilidad compartida.

Un modelo de fragilidad compartida puede ser considerado como un modelo mixto (efectos fijos y aleatorios) en análisis de sobrevida con variación de grupo (fragilidad) y variación individual descrita por la función de riesgo. En contraste, los modelos mixtos muestran un manejo más simétrico de estas dos fuentes de variación. Debido a observaciones censuradas el modelo de Cox y los modelos de fragilidad pertenecen a la clase de modelos lineales mixtos generalizados. Se supone que hay independencia entre las observaciones de diferentes "conglomerados". Si la varianza de la variable de fragilidad es cero, esto implica la independencia entre los tiempos hasta el evento para los "conglomerados", de lo contrario, existe una dependencia positiva entre los tiempos hasta el evento para aquellos.

[10] en general el conglomerado es la unidad a la que se le observan los tiempos hasta un evento determinado. Por ejemplo, en el seguimiento de la recurrencia de un problema genético en una familia, el cluster es la familia

CAPÍTULO 2

Un modelo de sobrevida para eventos recurrentes con evento terminal

2.1. Modelo de fragilidad compartida para eventos recurrentes y un evento terminal

En este capítulo se presenta la propuesta para modelar un proceso de eventos recurrentes con un evento terminal. El modelo se propone siguiendo a Liu et al. (2004).

En estudios médicos de tipo longitudinal, se puede observar la ocurrencia repetida de uno o varios eventos en un mismo individuo, situación que puede conducir a la muerte. En casos de hospitalizaciones sucesivas, casos de recaídas reiterativas por ejemplo, luego de la extirpación de un tumor, normalmente conllevan a la decaida del paciente y pueden llevarlo a la muerte. En la práctica, la ocurrencia del evento recurrente y del evento terminal no son independientes. La ocurrencia de ataques al corazón frecuentemente aumenta el riesgo de muerte. Esta dependencia debe tenerse en cuenta en el modelamiento conjunto de eventos recurrentes y evento terminal. En este modelo conjunto propuesto, la dependencia de los dos procesos, se modela mediante el acondicionamiento de un efecto aleatorio compartido (fragilidad compartida) que se incluye en ambas funciones de riesgo. Mediante este tipo de modelamiento conjunto, se tiene la posibilidad de medir el impacto de la ocurrencia de los eventos recurrentes en la ocurrencia del evento terminal.

De esta manera, un modelo de *fragilidad compartida* se propone para modelar conjuntamente los dos procesos eventos recurrentes y evento terminal, que se sospechan dependientes. Es un modelo conjunto semiparamétrico para las funciones de intensidad de tales eventos. Es conjunto a través de una fragilidad gama compartida. En este modelo se ajusta un parámetro, η, que modifica de manera exponencial el término de fragilidad del modelo del riesgo terminal, que permite determinar la dependencia (o independencia) del proceso del evento terminal del proceso del evento recurrente observado.

Con esta forma de disponer el coeficiente de fragilidad, se puede estimar un término que de alguna manera diferencie un impacto del término de fragilidad en el proceso del evento terminal, donde se supone que el desarrollo de este proceso puede impactarse de manera importante a medida que ocurre el evento recurrente. La posibilidad de tener una medida

del nivel de dependencia ($\nu > 0$) del proceso del evento terminal con respecto al proceso del evento recurrente, es el propósito por el cual se asume este tipo de modelamiento.

Dado el tiempo de censura C_i y el tiempo del evento terminal T_i, se escribe $\tau_i = min(C_i, T_i)$, como el tiempo de seguimiento y $\Delta_i = I(T_i \leq C_i)$, donde $I(.)$ es una función indicadora. Sea $X_i(t) = I(\tau_i \geq t)$ la indicadora que el sujeto está en riesgo. Se denota por $N_i^{T*}(t) = I(T_i \leq t)$ y $N_i^T(t) = I(\tau_i \leq t, \Delta_i = 1)$, el indicador de muerte (evento terminal) real y observado respectivamente, durante el tiempo t. Igualmente se define $N_i^{R*}(t)$ y $N_i^R(t)$ como el número real y observado de ocurrencias del evento recurrente para el individuo i, con $N_i^R(t) = N_i^{R*}(min(\tau_i, t))$. Sea $dN_i^{R*}(t) = N_i^{R*}\{(t + dt)-\} - N_i^{R*}(t-)$ cuando $dt \to 0$ y $dN_i^R(t) = I(\tau_i \geq t)dN_i^{R*}(t)$.

Se introduce heterogeneidad con covariables observadas z_i y fragilidad no observada ν_i, la cual mide el estado de riesgo del individuo relativo al evento recurrente y al evento terminal. La observación por sujeto i es $\mathbf{O_i}(t) \equiv \{X_i(u), N_i^R(u), N_i^T(u), 0 \leq u \leq t\}$ una copia independiente e identicamente distribuída (i.i.d) hasta el tiempo t de los datos completos observados $\mathbf{O} = \{\mathbf{O}(t), 0 \leq t \leq \tau\}$. Nótese que para cada individuo, $\mathbf{O_i}$ reúne la información de si el individuo se encuentra en riesgo al tiempo t por los dos procesos, el número de ocurrencias del evento recurrente que ha tenido hasta el tiempo t y la indicadora de riesgo de muerte (evento terminal) al tiempo t. $\mathbf{O}$ reúne la información del proceso conjunto.

Se define F_0 el σ-campo generado por (ν, z) y $F_t = \sigma\{F_0, \mathbf{O}(u), 0 \leq u \leq t\}$.

Los siguientes supuestos se hacen sobre los procesos subyacentes (el de recurrencias y el del evento terminal):

1. Los procesos de los eventos terminal, recurrentes y el proceso de censura tienen distribución continua, por lo cual no pueden suceder el mismo tiempo. Se adopta la convención de que el evento muerte (terminal) pasa primero en el intervalo $[t + dt)$. Se asume que la ocurrencia del evento terminal detiene la ocurrencia de cualquier otro evento recurrente y que $N_i^{R*}(t)$ es constante después de T_i.

2. $P(dN^T(t) = 1 \mid F_{t^-}) = X_i(t)d\Upsilon_i(t) \equiv X_i(t)\gamma_i(t)dt$. Donde $d\Upsilon i(t) = P(dN^{T*}(t) = 1 \mid z, \nu, T \geq t)$. Esto es, la probabilidad de que el evento muerte/terminal suceda, dadas las covariables observadas y el término de fragilidad es equivalente a la indicadora de que el individuo está en riesgo, multiplicada por la función estimada del riesgo de muerte en un momento infinitesimal.

3. $P(dN^R(t) = 1 \mid F_{t^-}, T \geq t) = X_i(t)d\Lambda_i(t) \equiv X_i(t)\lambda_i(t)dt$. Donde $d\Lambda_i(t) = P(dN_i^{R*}(t) = 1 \mid z_i, \nu_i, T_i \geq t)$. Igual a la interpretación del item anterior, esto supone que la probabilidad de que los eventos recurrentes sucedan para el individuo i al tiempo t, dadas las covariables observadas y el término de fragilidad es equivalente a que la indicadora de que el individuo está en riesgo, multiplicado por la función estimada del riesgo de los eventos recurrentes en un momento infinitesimal.

   Se hace notar que $P(dN^R(t) = 1 \mid F_{t^-}, T \leq t)$ no es estimable generalmente y es cero en esta forma, donde T marca un evento terminal. Los eventos $(T < t)$ y $(T \leq t)$ es casi seguro que son idénticos puesto que T tiene una distribución contínua.

4. La censura es no informativa. Es decir la censura no depende de ν.

5. $P(N^R(\tau > 1) > 0$, lo cual asegura que ν y η (coeficientes del término de fragilidad en el modelo del evento terminal) puedan ser identificados.

2.1.1. Estimación de los coeficientes del modelo

Como en Kalbfleisch and Prentice (2002), la verosimilitud completa puede ser escrita como el producto integral:

$$\boldsymbol{L} = L(F_0)L(\mathbf{O} \mid F_0) \tag{2.1}$$

Ignorando la contribución de la censura independiente o no informativa,

$$L(\mathbf{O_i} \mid F_0) = P_0^{\infty} L(F_{t^-+dt} \mid F_{t^-}) \propto P_0^{\infty} L(dN_i^R(t), dN_i^T(t) \mid F_{t^-}) \tag{2.2}$$

Similar al desarrollo para encontrar la verosimilitud para riesgos en competencia:

$$\begin{aligned} &L(dN_i^R(t), dN_i^T(t) \mid F_{t^-}) \\ &\qquad = [X_i(t)d\Upsilon_i(t)]^{dN_i^T(t)}[1 - X_i(t)d\Upsilon_i(t)]^{1-dN_i^T(t)} \\ &\qquad\quad \times \left\{[X_i(t)d\Lambda_i(t)]^{N_i^R(t)}[1 - X_i(t)d\Lambda_i(t)]^{1-dN_i^R(t)}\right\}^{1-dN_i^T(t)} \end{aligned} \tag{2.3}$$

Se hace notar que se adopta el hecho de que $0^0 = 1$. Para eventos en tiempo continuo 2.3 es equivalente a

$$\begin{aligned} &[X_i(t)d\Upsilon_i(t)]^{dN_i^T(t)}[1 - X_i(t)d\Upsilon_i(t)]^{1-dN_i^T(t)} \\ &\qquad\qquad \times [X_i(t)d\Lambda_i(t)]^{N_i^R(t)}[1 - X_i(t)d\Lambda_i(t)]^{1-dN_i^R}(t) \end{aligned} \tag{2.4}$$

o

$$P_0^{\infty} L(dN_i^R(t), dN_i^T(t)|F_{t^-}) = P_0^{\infty} L(dN_i^R(t)|F_{t^-}, D > t) P_0^{\infty} L(dN_i^T(t)|F_{t^-}) \tag{2.5}$$

De esta manera, ampliando el modelo de Huang y Wolfe (2002), los modelos para ambos procesos quedan de la siguiente manera:

$$\lambda_i(t) = \nu_i \exp(\beta^{'} z_i)\lambda_0(t) \tag{2.6}$$

$$\gamma_i(t) = \nu_i^{\eta} \exp(\alpha^{'} z_i)\gamma_0(t) \tag{2.7}$$

Para los procesos de los eventos recurrentes y terminal respectivamente.

La presencia del parámetro de fragilidad común ν, debilita el supuesto usual de censura no informativa de proceso de eventos recurrentes en el proceso del evento terminal. Se adopta la función de fragilidad gama $f_\theta(\cdot)$ con media 1 y varianza θ. La media es 1 para evitar el problema de no identificabilidad, el cual podria surgir si se multiplica y se divide el término de fragilidad y el riesgo base por la misma constante. La escogencia del tipo de distribución que debería tener el término de fragilidad se debe en gran medida a que sea una función que sea al menos de soporte positivo. Ya que el modelo de fragilidad es un modelo multiplicativo y la función de riesgo por definición es de soporte positivo, la función de distribución del término de fragilidad se propone también de soporte positivo. Cuando para el ajuste de este tipo de modelos se usan programas computacionales como el PROC NLMIXED de SAS, que requieren que los efectos aleatorios tengan una distribución normal, se propone hacer una transformación en la programación para obtener que la distribución de los efectos aleatorios sea no normal (Nelson et al. (2006)). Puesto que en este trabajo se usa el PROC NLMIXED para el ajuste del modelo, en el apéndice 2 se muestra la transformación para lograr que la distribución de los efectos aleatorios sea gama con media 1 y varianza θ.

El modelo utilizado para el manejo de las recurrencias es el propuesto por Andersen y Gill (1982), donde cada sujeto es tratado como un proceso contador con sucesos múltiples y con tiempos inter-ocurrencias (los tiempos entre evento y evento, incluída la ocurrencia del evento terminal) independientes, dada la historia de todas la variables observadas hasta el tiempo de presentación de los eventos.

Cuando $\eta = 0$, $\gamma_i(t)$ no depende de ν_i y es no informativa para la tasa de eventos recurrentes $\lambda_i(t)$. $\theta = 0$ implica que los términos de fraglidad ν_i son idénticamente 1, es decir, la tasa de eventos tanto de los recurrentes como el terminal, se explica únicamente por z_i.

Si η y θ son significativos, se concluye la dependencia del proceso del evento terminal con respecto al proceso de las recurrencias y la heterocedastidad existente, ya sea por la variabilidad en la que se pueden presentar los riesgos entre los individuos, o por que el riesgo para los dos procesos no es explicado completamente por las covariables tenidas en cuenta en el modelamiento.

Si η es significativo y θ no lo es, se concluye la dependencia entre los procesos, pero no se valida la hipótesis de heterocedasticidad existente. Se puede asumir que los riesgos entre los individuos pueden ser proporcionales y que, las covariables tenidas en cuenta en el modelamiento para cada uno de los procesos, explican completamente la tasa de riesgo tanto a la ocurrencia de los eventos recurrentes como del terminal.

Si η no es sigificativo y θ sí, se concluye la no dependencia entre los procesos, y se puede concluir la heterocedasticidad sospechada, es decir que el riesgo a la ocurrencia de las recurrencias como al evento terminal sucede de manera no proporcional (desigual) entre los individuos, y que las covariables tenidas en cuenta en el modelamiento para los dos modelos (el de las recurrencias y el terminal) no son suficientes para explicar el riesgo a la ocurrencia de los eventos.

Dado t_{ij} el j-ésimo tiempo del evento recurrente para el i-ésimo sujeto, y sea δ_{ij} la indicadora de la ocurrencia del evento recurrente al tiempo t_{ij}. Sea $\underline{\mathbf{t}}_i$ el tiempo de observación total para el individuo i. El primer factor de la verosimilitud (2.5) es

$$\exp\left[\int_0^\infty X_i(t)\nu_i \exp(\beta^{'} z_i) d\Lambda_0(t)\right] \times \prod_j \left[\nu_i \exp(\beta^{'} z_i) d\Lambda_0(t_{ij})\right]^{\delta_{ij}} \tag{2.8}$$

Similarmente,(2.8) sugiere que el segundo término de la verosimilitud (2.5) es proporcional a

$$\exp\left[\int_0^\infty X_i(t)\nu_i^\eta \exp(\alpha^{'} z_i) d\Upsilon_0(t)\right] \times \left[\nu_i^\eta \exp(\alpha^{'} z_i) d\Upsilon_0(\mathbf{t}_i)\right]^{\Delta_i} \tag{2.9}$$

Con esto, $L(\mathbf{O_i} \mid F_0)$ puede escribirse como la mutiplicación de (2.8) y (2.9), que son las verosimilitudes de los dos procesos, el de las recurrencias y el del evento terminal. Así se tiene entonces la verosimilud conjunta, en donde el término de fragilidad se incluye como uno solo para ambos procesos, garantizando de esta manera el supuesto de heterogeneidad conjunta para los procesos. Además se asume el término η que indica incidencia del proceso de las recurrencias sobre el terminal. La verosimilitud completa para $\{(\mathbf{O_i}, \nu_i), i = 1, ..., n\}$ es

$$\begin{aligned} l = \log \prod_{i=1}^n L(\mathbf{O_i}, \nu_i \mid z_i) &= \log \left[\prod_{i=1}^n L(\mathbf{O_i}, \nu_i \mid z_i) f_\theta(\nu_i)\right] \\ &= \sum_{i=1}^n \left[\sum_j \left[\log(\nu_i + \beta^{'} z_i + \log d\Lambda_0(t_{ij})\right] - \int_0^\infty X_i(t)\nu_i \exp(\beta^{'} z_i) d\Lambda_0(t)\right] \\ &\quad + \sum_{i=1}^n \left[\Delta_i \left[\eta \log \nu_i + \alpha^{'} z_i + \log d\Upsilon_0(\mathbf{t}_i)\right] - \int_0^\infty \nu_i^\eta \exp(\alpha^{'} z_i) d\Upsilon_0(t)\right] \\ &\quad + \sum_{i=1}^n \log f_\theta(\nu_i). \end{aligned} \tag{2.10}$$

Si se desea realizar la estimación de los parámetros utilizando los tiempos inter-ocurrencias, la expresión de la verosimilitud es la misma excepto que t_{ij} debe ser reemplazado por $g_{ij} = t_{ij} - t_{ij-1}$(Rondeau(2007)). En la aplicación realizada en este trabajo se utilizaron los tiempos inter-ocurrencias.

Para la estimación de los parámetros del modelo del proceso de recurrencias, ninguna estructura de dependencia entre los tiempos hasta cada ocurrencia del evento recurrente es impuesta, de la misma manera como lo propone el modelo marginal de Wei, Lin y Weissfeld (sección 1.2.3). Los parámetros del modelo son estimados directamente de la función de verosimilitud.

2.1.2. Método para la estimación de los parámetros

La ecuación (2.10) suministra la verosimilitud para los datos completos con términos de fragilidad conocidos lo cual es más fácil de maximizar que la verosimilitud de los datos

observados. Esto hace que el algoritmo EM sea una selección natural para la estimación de los parámetros. En el paso E, puesto que no hay una forma cerrada para la densidad de $f(\nu_i|\mathbf{O_i})$, el algoritmo Metropolis Hasting se usa para generar M números aleatorios $\nu_i^m (m = 1, ..., M)$ para la estimación de la esperanza de los estadísticos suficientes involucrando las fragilidades.

En el paso M, las estimaciones de los parámetros se obtienen maximizando la expresión (2.10) como si los estadísticos de la fragilidad fueran conocidos. Los componentes de las derivadas parciales para β y $\lambda_0(.)$ son

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left[\sum_j \delta_{ij} z_i(t_{ij}) - \int_0^\infty X_i(t) z_i(t) \hat{E}(\nu_i|\mathbf{O_i}) \exp(\beta' z_i(t)) d\Lambda_0(t)) \right] \tag{2.11}$$

$$\frac{\partial l}{\partial \lambda_0(t_{ij})} = \frac{\delta_{ij}}{\lambda_0(t_{ij})} - \sum_{k=1}^{n} X_k(t_{ij}) \hat{E}(\nu_i|\mathbf{O_i}) \exp(\beta' z_k(t_{ij})) \tag{2.12}$$

La estimación de la funcion de riesgo base para el proceso de eventos recurrentes se hace mediante la estimación de Breslow (1975)

$$\hat{\lambda}_0(t_{ij}) = \frac{\delta_{ij}}{\sum_k X_k(t_{ij}) \hat{E}(\nu_k k|\mathbf{O_k}) \exp(\beta' z_k(t_{ij}))} \tag{2.13}$$

Con esto β puede derivarse sustituyendo (2.13) en (2.11). La segunda derivada parcial para el coeficiente de covariables β es

$$\frac{\partial^2 l}{\beta^2} = -\sum_{k=1}^{n} \int_0^\infty X_i(t) z_i(t) z_i(t)' \hat{E}(\nu_i|\mathbf{O_i}) \times \exp(\beta' z_i(t)) d\Lambda_0(t)) \tag{2.14}$$

Puesto que el algoritmo EM no proporciona directamente la matriz de información para la verosimilitud de los datos observados, se usa la formula de Louis(1982), para obtenerla. Sea $\rho = (\beta, \alpha, \eta, \theta, \lambda_0, \gamma_0)$. La matriz de información observada $I(\hat{\rho})$ está dada por

$$I(\hat{\rho}) = -\hat{E}\left[\frac{\partial^2 l}{\partial \rho \partial \rho'} \mid \mathbf{O}, \hat{\rho} \right] - \hat{E}\left[\frac{\partial l}{\rho} \frac{\partial l}{\rho'} \mid \mathbf{O}, \hat{\rho} \right] + \hat{E}\left[\frac{\partial l}{\rho} \mid \mathbf{O}, \hat{\rho} \right] \hat{E}\left[\frac{\partial l}{\rho'} \mid \mathbf{O}, \hat{\rho} \right] \tag{2.15}$$

Todos estos términos se evalúan en la última iteración del algoritmo EM cuando el último término se convierte en cero para la estimación maximo-verosímil de ρ. La primeras dos esperanzas se pueden calcular mediante el promedio de los términos correspondientes que involucran los valores obtenidos mediante el algoritmo Metropolis Hasting.

2.2. Censura

Se considera la censura para estos dos modelos a derecha, esto es, un individuo es censurado cuando $\Delta_i = 0$, es decir, cuando no se conoce el tiempo de ocurrencia del evento terminal antes del tiempo de finalización del estudio. En este modelamiento conjunto, la ocurrencia del evento terminal da fin a la ocurrencia de los eventos recurrentes, asi que en el modelo de eventos recurrentes la censura está condicionada a la censura del evento terminal. En este sentido los dos modelos, el del proceso terminal y el del proceso de las recurrencias no son independientes, es decir, son informativas la una de la otra.

CAPÍTULO 3

Aplicación

3.1. Modelo de sobrevida para eventos recurrentes con evento terminal en deserción de clientes

3.1.1. Introducción

En la industria de las telecomunicaciones, dada la alta competitividad por el número de operadores disponibles en el mercado, se presenta un fenómeno que es muy difícil de gestionar y además de muy alto impacto financiero para las compañías prestadoras de servicios de telecomunicaciones (en adelante telcos), denominado en el argot de mercadeo deserción de clientes (en inglés *churn*). Esta situación tiene implicaciones negativas las cuales fundamentalmente se pueden resumir en : 1) la pérdida de participación en el mercado tanto en valor como en clientes y la más importante 2) la pérdida o el deterioro de los ingresos en el caso de no tener una dinámica de recuperación de clientes lo suficientemente eficiente. En términos del manejo estratégico para el nombre e imagen de la compañía y del manejo financiero de la misma, cualquiera de las dos implicaciones es grave. No obstante, como ésta es una realidad que no se puede evitar, normalmente se determina un umbral que define el nivel de aceptación de pérdida de clientes por un período de tiempo definido. Dada esta situación, las telcos han creado en su estructura organizacional áreas dedicadas exclusivamente al control de este umbral, lo que determina dispendiosos y costosos procesos de gestión de clientes con el propósito de impedir que éstos decidan no ser clientes de la compañía en un momento dado. Procesos exactamente denominados retención, blindaje y fidelización[1] son los que más comúnmente se realizan para impedir que los clientes deseen dejar de ser clientes de la compañía. la deserción de clientes sucede sin distingo del tipo de cliente y es más alta la deserción en algunos tipo de cientes dependiendo del segmento al que pertenecen.

Para determinar cómo se gestionan los diferentes tipos de clientes de las telcos, estas tienen segmentados sus clientes. Los segmentos naturales son los de personas naturales y empresas. Éstos se subdividen en otros con el fin de poder generar las estrategias adecuadas

[1]La retención es una reacción al momento en que el cliente solicita que le retiren el servicio. Lo que se hace generalmente es mejorarle las condiciones de servicio al cliente. El blindaje consiste en que se le ofrece al cliente por el mismo precio, más o mejores servicios; y la fidelización consiste entre otras cosas en sorprender al cliente con algún beneficio que él normalmente no estaba esperando.

para cada uno de ellos. Los subsegmentos se definen normalmente según el valor de renta mensual de cliente ó también por el valor que el cliente está dispuesto a pagar por los servicios de telecomunicaciones que requiere (la renta mensual del cliente con una compañía de telecomunicaciones puede no ser lo que éste está dispuesto a pagar o está pagando por estos servicios ya que, puede tener servicios contratados con otros operadores). Este tipo de segmentación es muy eficiente porque lleva a entender cuáles son los clientes que son más valiosos para la compañía. El involucramiento del tiempo de vida esperado de los clientes también es una herramienta que permite determinar el valor de un cliente.

La información utilizada para este análisis corresponde a un segmento de empresas que está subdividido en las grandes empresas (estratégicas, grandes e intermedias), y las mipymes (medianas, pequeñas y micro empresas). Las grandes empresas son los clientes de las telcos que más generan ingreso recurrente. Son los clientes más valiosos para la compañía. Un cliente grande promedio con una solución de servicios de telecomunicaciones muy completa puede llegar a facturar hasta 100 veces y más de lo que factura una persona que tenga el promedio de los servicios de telecomunicaciones para el segmento personas. Es por esto que este tipo de clientes requieren de mucha atención y esfuerzos para evitar su deserción.

En la industria de las telecomunicaciones ha ido creciendo la cantidad de servicios que puede usar un cliente para mejorar sus niveles de comunicación. A medida que transcurre el tiempo, esta industria crece en función de los grandes avances tecnológicos. No es desconocido, por ejemplo, el surgimiento de un nuevo servicio que está en auge, denominado *Cloud Computing* [2], que más allá de ser "un servicio"es una gran cantidad de servicios reunidos en este término. Una empresa puede tener una gran cantidad de servicios para dar comodidad a sus clientes. Es así, que uno de los retos que tienen los encargados de las tecnologías de las información (IT) en las grandes empresas es garantizar que todos los servicios de telecomunicaciones estén en "alta", es decir que estén disponibles todo el tiempo. Esta necesidad varía según el sector económico. Por ejemplo para el sector financiero la disponibilidad debe ser muy cercana al 100 %. Para garantizar esto los encargados de IT recurren a tener más de un operador para mantener estas condiciones. Los servicios que adquieren los clientes (de cualquier segmento) son instalados mediante un "enlace"[3]que determina el tipo de productos/servicios con los que el cliente cuenta. Una empresa grande puede llegar a tener más de 100 enlaces conectados. A la colección de servicios de telecomunicaciones que adquiere una empresa se le denomina "solución". Esta solución debe ser óptima tanto en niveles de servico, operatividad y atención, para que los clientes mantengan el "paquete"total de la solución.

Cuando un cliente grande toma una solución de servicios con un operador de telecomunicaciones la implementación de ésta en cierto sentido es bastante compleja, por lo que en algunos casos cambiar de operador dada una solución ya instalada, suele ser desgastante para el operador de telco y riesgoso para el mismo cliente. Dadas estas premisas y las altas exigencias de servicio que exigen estos clientes, es común el hecho de que en un momento de insatisfacción con los servicios prestados y en aras de mitigar el riesgo de la migración y desinstalación de toda la solución, los clientes opten antes de hacer una migración total de los servicios a otro operador, ir desinstalando los enlaces asociados a los servicios de

[2]Servicio mayormente entendido como la posibilidad de tener acceso a muchas aplicaciones de software que normalmente se adquieren por demanda - también se entiende como servicios en la nube - entre otra gama de servicios asociados a la posibilidad tener la información en un servidor dispuesto por el operador con diversos niveles de seguridad y acceso

[3]Un enlace es la conexión que se le hace al cliente a la red de un operador

manera paulatina. Esto puede provocar que se pierda el cliente totalmente o lo inevitable, la pérdida del ingreso recurrente asociado a esa conexión de servicio. A este fenómeno se le denomina "***deserción pasiva***".

Conocido este precedente, el objetivo fundamental es modelar estos dos fenómenos. Se centra el interés en evitar el hecho definitivo de pérdida de clientes y poder controlar la pérdida del ingreso por las desconexiones previas al deceso. Para el logro de este objetivo se consideraron varias posibilidades de modelamiento del riesgo: inicialmente se consideró un modelo de Cox con covariables las mismas usadas para el modelamiento de esta última propuesta presentada en esta tesis. Frente a la evidencia de un ajuste pobre de este modelo y el conocimiento de que existía el hecho de la desconexión de enlaces que conllevaba a pérdida de ingreso y de los clientes, se determinó tomar una vía de modelamiento para tiempos de sobrevida multivariados. Frente a este nuevo panorama, en segunda instancia se consideró un modelo de riesgos en competencia bajo la expectativa de determinar niveles de riesgo por los diferentes factores que pudieran causar la pérdida del cliente (precio, insatisfacción con el servicio, mejor oferta de otro proveedor, problemas con las condiciones iniciales del contrato, mala atención por parte del área de servicio al cliente, mala atención postventa, entre otros) y agregando una covariable que diera cuenta del número de desconexiones previas por cliente. Este tipo de modelamiento se descartó por la poca y deficiente información existente en los sistemas, acerca de la causa real de desinstalación total del cliente. Normalmente como este es un dato del cliente que ingresan al sistema los técnicos de desinstalaciones, no es obligatorio por no ser la función princial de aquellos, por lo que, la información es muy pobre y casi nula. Finalmente mediante el análisis de factores a incluir en el modelo que pudieran dar mejor acercamiento para el manejo eficiente de una gestión de cliente se determinó incluir con contundencia el proceso subyacente a la pérdida de clientes -el de la deserción pasiva -, por lo que después de varias alternativas de modelamiento se optó tener en cuenta el modelamiento de múltiples eventos recurrentes con evento terminal, con la idea de tener en cuenta las recurrentes desinstalaciones de enlaces por cliente.

En este capítulo se presenta el modelamiento de eventos recurrentes con evento terminal para la estimación del riesgo de pérdida de clientes en el sector de las telecomunicaciones, teniendo en cuenta un proceso subyacente de desconexión previa de enlaces que puede impactar el hecho de la pérdida de un cliente, pero que a la vez puede dar información -mediante su adecuado modelamiento- para prevenir este riesgo. La idea fundamental consiste en encontrar argumentos estadísticos para generar gestiones tendientes a mitigar la pérdida de clientes en el segmento de las grandes empresas, dado el precedente conocido de la deserción pasiva.

3.1.2. Definición de las recurrencias

Las recurrencias se definen como las desinstalaciones sucesivas de los enlaces asociados a los servicios instalados a cada cliente. Los tiempos de las recurrencias son tomados como los tiempos de espera entre una y otra. Este mecanismo es apropiado para cuando la predicción del tiempo para el próximo evento es de interés, cuando el individuo (en este caso el cliente -la empresa-) sufre algún tipo de renovación (Cook & Lawless (2006), Sección 1) o cuando los eventos son relativamente poco frecuentes. Esta mecánica de tomar los tiempos entre desintalaciones es adecuado para el caso en desarrollo para esta tesis, puesto que normalmente luego de la desinstalación de algún enlace el cliente puede que tenga algún beneficio en otro servicio con el propósito de hacer una gestión de retención.

Es claro que esto denota un efecto de renovación del cliente. Ahora, desde el punto de vista de predicción para efectos de gestión del cliente es importante poder tener un dato del tiempo al cual estaría este cliente propenso a tener una nueva desintalación de un servicio.

Un hecho que no se debe desconocer es que los clientes no siempre instalan todos sus servicios en el momento en que ingresan a la compañía (momento que para el análisis es el tiempo ($t = 0$)). Hay servicios que pudieron ser instalados luego de que el cliente ya tenía instalados algunos servicios. Para un cliente, un gráfico de instalaciones y desintalaciones de los enlaces asociados a sus servicios es el que se presenta en la figura 3.1.

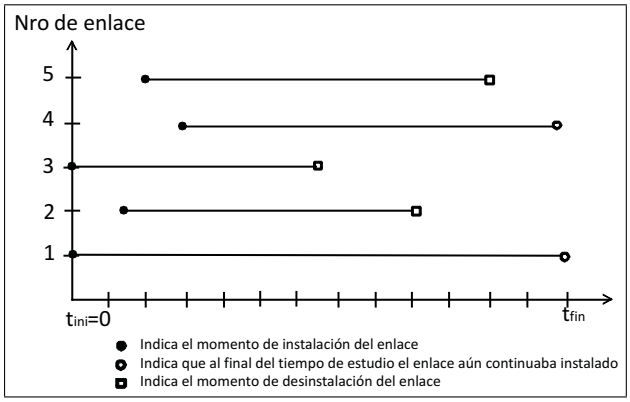

FIGURA 3.1. Instalación y desinstalación de enlaces por cliente .

Esta figura representa un proceso de instalaciones y desinstalaciones de un cliente "tipo"incluido en el estudio. Este cliente desde el inicio de su vida en la compañía $t_{ini} = 0$ y hasta el final del tiempo de estudio t_{fin} le fueron instalados 5 enlaces: los enlaces 1 y 3, se instalaron en el momento del ingreso del cliente a la compañía, mientras que los enlaces 1, 4 y 5 se instalaron luego de su ingreso. Los enlaces 2, 3 y 5, fueron desinstalados antes de culminar el tiempo de estudio, mientras que los enlaces 2 y 3 aún continuaban instalados al final del tiempo de estudio. Valga aclarar que si el cliente desinstala todos los servicios definitivamente antes de t_{fin}, entonces T_i=t_{fin}, donde T_i es el tiempo de ocurrencia del evento terminal para el cliente i.

El modelamiento se puede realizar con los tiempos calendario o con los tiempos inter-ocurrencias. Para este trabajo se utilizan éstos últimos. La construcción de estos tiempos (también denominados tiempos de espera) se realiza haciendo como tiempo cero el momento de ingreso del cliente a la compañía y desde ese momento se contabiliza el tiempo hasta la primera desinstalación(t_1), luego desde la primera hasta la segunda se vuelve a tomar el tiempo para la segunda recurrencia (t_2) y así sucesivamente, como se muestra en la figura 3.2.

Para este cliente $N(t) = 3$ que es el número acumulado de eventos (desinstalaciones) ocurridos en el intervalo de estudio para este cliente $(0, t_{fin})$.

3.1.3. Recurrencias con evento terminal

El evento terminal lo define la desinstalación total del cliente. Cuando sucede el evento terminal no sucede ninguna otra recurrencia- es decir, este evento es absorbente-. Se debe evidenciar que el cliente está definitivamente fuera de la compañía para no generar sesgos

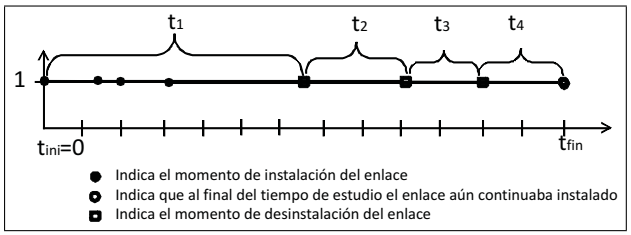

FIGURA 3.2. Construcción de los tiempos de espera por individuo - cliente.

en la estimación del riesgo. La pérdida del cliente se debe a que se retira para contratar los servicios con otro operador, es decir abandona la compañía totalmente. Es común en las empresas grandes que se desconectan por razones tan sencillas como el traslado de sus sedes o porque piensan mejorar su solución de servicios, por ejemplo, desconectan enlaces de muy bajas velocidades para instalar de altas velocidades. Vuelven a los pocos meses presentando una recurrencia de ingreso a la compañía. Este tipo de ausencias no son tomadas como clientes perdidos, así se hayan desinstalado por períodos de tiempo grandes ($>$ 6 meses).

Dada la historia de vida del cliente en la empresa con respecto a un proceso de deserción (cuando empieza a desinstalar los servicios), se puede considerar que los tiempos asociados a este proceso estén relacionados, es decir, que el hecho de que sucedan en diferentes momentos recurrentemente desinstalaciones de los servicios para el cliente, puede generar un tiempo de desintalación total consistente con el proceso de desintalaciones: a mayor número de desinstalaciones en períodos consecutivos de tiempo es de esperar que este cliente tenga mayor riesgo de ser desinstalado. Este hecho indica la necesidad de tener en cuenta un término de fragilidad que permita determinar el nivel de asociación entre el proceso de las recurrencias y el proceso de deserción (muerte) total del cliente.

3.1.4. Modelo conjunto de eventos recurrentes y evento terminal

Definir procesos de gestión adecuados para los clientes implica tener acciones más preventivas que reactivas. El propósito del modelamiento finalmente es poder tener una medida de riesgo que permita predecir el momento o riesgo de una desinstalación y determinar cuál es la tasa de riesgo de pérdida de clientes, de tal manera que sea factible en un momento dado pronosticar, con altos niveles de confianza, el valor de los clientes en la industria de las telecomunicaciones y así generar el proceso de administración de clientes bastante efectivo. Además de, determinar los niveles de riesgo asociados con las covariables incluidas en el modelamiento.

Las covariables seleccionadas son tales que puedan ser accionables de una manera tangible. Un ejemplo de una variable accionable es la segmentación del cliente (que define entre otras cosas, el tipo de atención comercial que se le da al cliente) ya que, si esta variable resulta tener un impacto importante negativo en alguno de los riesgos estimados (tanto el de las recurrencias de desinstalaciones como en la pérdida total del cliente), pueden realizarse las respectivas modificaciones en la fuerza de ventas y en otros departamentos de la compañía que manejan esta variable para su gestión, para propender a que este impacto disminuya.

Las covariables elegidas según la propuesta teórica son incluídas en el modelo conjunto, lo cual indica que se pueden estimar coeficientes que permitan determinar cómo varían los niveles de riesgo tanto para las recurrencias como para el riesgo de desinstalación total (evento terminal) por cada una de las variables de interés. Más adelante se describen las covariables incluídas en el modelamiento.

El modelamiento se propone de manera conjunta ajustando modelos para los dos eventos por separado e incluyendo un término de fragilidad compartido. Tener un modelo conjunto, donde el término de fragilidad es compartido permite, además de tener las interpretaciones debidas a la estimación de los parámetros para cada modelo y del término de fragilidad, entender cuál proceso evoluciona más rapido en promedio.

El cálculo del cociente $\frac{\lambda_i(t)}{\gamma_i(t)}$ para el grupo de individuos en estudio puede dar nociones acerca de cuál proceso es en un momento dado más acelerado. Aunque una estrategia para no perder clientes gestionándolos desde el conocimiento del comportamiento y niveles de riesgo obtenido mediante el modelo de las recurrencias puede ser muy exitosa, el encontrar un momento en donde se acelera el proceso de riesgo de pérdida de clientes (deserción - muerte), es decir poder determinar si el riesgo del evento terminal avanza mas rápido que el riesgo del evento de las recurrencias o al contrario, es un dato muy útil para la compañía, ya que permite determinar de qué nivel puede ser la pérdida del ingreso proyectada a un tiempo t en el caso de no desarrollar adecuadas estrategias de gestión de la deserción de clientes.

De esta manera, y de acuerdo con las covariables consideradas en el modelo, el modelamiento conjunto corresponde entonces al ajuste de los dos modelos:

$$\lambda_i(t) = \nu_i \exp(\beta_1 * X_1 + \beta_2 * X_2 + .. + \beta_p * X_p)\lambda_0(t) \tag{3.1}$$

$$\gamma_i(t) = \nu_i^{\eta} \exp(\alpha_1 * X_1 + \alpha_2 * X_2 + ... + \alpha_p * X_p)\gamma_0(t) \tag{3.2}$$

Para la estimación del riesgo asociado con las recurrencias y al evento terminal respectivamente y con p, el número de covariables consideradas para el modelamiento.

3.1.5. Consideraciones técnicas acerca de los datos

La información para el analisis corresponde al tiempo de vida de los servicios de los clientes en una empresa de telecomunicaciones. Un cliente es una entidad comercial que paga a la empresa por uno o más servicios de telecomunicaciones. Los servicios que puede prestar la empresa son, globalmente, siete : servicios de Voz, de datos, internet, e-business, servicios administrados, servicios de red y otros servicios (que más que un servicio es una categoría de servicios que se clasifican aquí).

Un cliente puede ingresar a la compañía por uno o más servicios -enlaces-, igualmente, puede desinstalar, uno o más servicios hasta que decide lo mismo para todos los servicios. El evento de falla sucede cuando el cliente desintala los enlaces asociados a cada uno de los servicios que tiene conectados y las recurrencias se deben a que el cliente sufre

desinstalaciones consecutivas de estos enlaces en todos los servicios hasta que se desinstala completamente.

La censura que se tiene es a derecha y de tipo I, es decir, el tiempo de un cliente es censurado cuando al final del período de estudio aún no ha desinstalado ningún servicio. El lapso de tiempo del presente estudio está determinado por las fechas de ingreso de la base de clientes obtenida al momento de extracción de la información. Esto es: se tienen en cuenta para el estudio todos los clientes tanto los actuales como los desconectados. El tiempo de inicio del estudio lo marcó el cliente con mayor antigüedad en el sistema. De esta manera, si $\underline{f}_i$ es la fecha de ingreso del cliente i a la compañía, la fecha de inicio para determinar el tiempo del estudio es el $min = \underline{f}_i$. El cliente con mayor antigüedad en la base de estudio tiene $11,5$ años, que concuerda con el tiempo de seguimiento o tiempo del estudio. La fecha de extracción de la información de los sistemas para el análisis fue el 31/08/2011.

El tiempo de sobrevida de cada individuo (cliente) es el mínimo entre dos tiempos: el de falla del evento terminal y el de censura.

Además se tiene en cuenta para hacer el modelamiento covariables concernientes al *tipo de atención comercial* que se le da al cliente, el *segmento estratégico del cliente*, la *ciudad de la sede principal* del cliente, la *pertenencia del mismo a un grupo de empresas*, el *número de servicios del cliente* y el *número de quejas* interpuestas a la empresa de parte del cliente por problemas técnicos con los servicios conectados hasta el día en que se extrajo la información.

3.1.6. Descripción de la información

Definiciones

- *tinicio*: Es una variable de tiempo representada en días, que identifica el inicio del período entre la ocurrencia de un evento y otro. El primer registro de *tinicio* para cualquier individuo en los datos es igual a cero.
- *tfin*: Es una variable de tiempo representada en días, que identifica el fin del período entre la ocurrencia de un evento y otro.
- *Recurrencia*: Un recurrencia es la desinstalación repetida de servicios o enlaces en un cliente, en momentos secuenciales del tiempo de vida de tal cliente en la empresa de telecomunicaciones.
- *evento*: El evento a estudiar para el caso del proceso de recurrencias, es la desconexión o desintalación de uno o más servicios a un cliente. Para el caso del proceso del evento terminal, es la desinstalación total del cliente (implica que ya el cliente no tiene ningún tipo de relación comercial con la empresa de telecomunicaciones).
- *Cliente desinstalado totalmente*: Es el estado de cliente que sucede luego de la desinstalación (o desconexión) total del cliente, es decir,cuando éste ya no cuenta con ninguno de los servicios que le presta la empresa de telecomunicaciones porque ya han sido desconectados todos los servicios.

TABLA 3.1. Fracción de los datos.

id	fecha_inicio	fecha_fin	tinicio	tfin	enum	evento	terminal
1	07/07/2005	11/08/2008	0	1940	1	0	1
2	12/04/2006	11/08/2008	0	852	1	1	0
2	11/08/2008	08/06/2009	852	1153	2	1	0
2	08/06/2009	04/11/2009	1153	1302	3	1	0
2	04/11/2009	27/10/2010	1302	1659	4	0	1
3	28/03/2005	12/07/2010	0	1932	1	1	0
3	12/07/2010	11/06/2011	1932	1999	2	0	1
4	07/07/2005	09/05/2006	0	302	1	1	0
4	09/05/2006	11/08/2008	302	1114	2	1	0
4	11/08/2008	04/11/2009	1114	1557	3	1	0
4	04/11/2009	31/08/2011	1557	2214	4	0	0
5	07/03/2006	31/08/2011	0	1995	1	0	1

TABLA 3.2. Fracción de los datos (continuación de 3.1)

id	Var Rta	Ciudad	segmento	Grup Obj	Nro Serv	Nro quejas
1	1940	1	1	1	1	2
2	852	2	2	2	2	2
2	301	2	2	2	2	2
2	149	2	2	2	2	2
2	357	2	2	2	2	2
3	1932	3	3	2	1	3
3	67	3	3	2	1	3
4	302	4	1	2	2	1
4	812	4	1	2	2	1
4	443	4	1	2	2	1
4	657	4	1	2	2	1
5	1995	1	3	1	1	2

- *Censura*: Es un estado de cliente determinado para el modelamiento, que determina si el cliente presentó o no el evento de interés en el tiempo de estudio. Se trata según lo descrito en la sección 2.5.

Una fracción de la información disponible a analizar se presenta en la tabla 3.1 y 3.2.

Cada línea corresponde a la información de las desinstalaciones de los enlaces del cliente en el período de estudio. En esta tabla se presenta la información para 5 clientes. Para recordar y ampliar lo mencionado en la introducción acerca del concepto de enlaces, si un cliente contrata con la compañía de telecomunicaciones los servicios de internet, datos y voz, puede tener bajo el servicio de internet 1 enlace de 1 MB , otro de 4 MB y otro de 10 MB; bajo el servicio de datos puede tener 2 enlaces: uno de 2000 MB y otro de 3000 MB y bajo el servicio de voz puede tener un enlace para un E1 (donde se conectan varias líneas) y 4 líneas telefónicas, las cuales cada una cuenta también con un enlace. En este sentido este cliente tiene 10 enlaces asociados a los servicios de telecomunicaciones con los que cuenta.

De esta manera en las tablas 3.1 y 3.2, se presenta una fracción de los datos. Cada fila muestra la información de ocurrencia de los eventos de interés. En este sentido varias filas muestran la información de un mismo cliente. La primera fila de las tablas , muestran la información recolectada para la empresa con $Id = 1$. A esta empresa, en el tiempo de estudio solamente le ocurrió un evento: la desintalación total (evento terminal). Esto quiere decir que independientemente del número de servicios que tenía, el cliente se desinstaló totalmente en un solo momento (el 11/08/2008).

Para este mismo cliente, en la columna 2, (*fecha_inicio*), y en la columna 3, (*fecha_fin*) se muestran las fechas en que se instaló por primera vez (07/07/2005) y la fecha en que se desinstaló completamente (11/08/2008); en la columna 4 se muestra el tiempo de inicio de observación del cliente desde que se ingresó a la compañía (*tinicio*=0), en la columna 5, se observa el tiempo de la desintalación total (*tfin*=1940), a los 1940 días; en la columna 6 se encuentra el contador de número de eventos resgistrados por cliente (*enum*=1), para este cliente esta variabe vale 1 ya que solo se le observó en el tiepo de estudio el evento terminal; la variable indicadora de la ocurrencia del evento desinstalación de un servicio (*evento*=0), se observa en la columna 7, que es igual a cero ya que este evento no se registra como recurrencia sino como terminal; la variable indicadora de la ocurrencia del evento terminal -la pérdida del cliente- se ve en la columna 8 (*terminal*=1), que es igual a uno por observarse el evento terminal en este cliente (desinstalación total); la variable respuesta (*Var Rta*=1940)-columna 1, tabla 3.2-, que es la cantidad de días que permaneció el cliente con la empresa (con servicios instalados); y por último, en la tabla 3.2 se observan las 5 covariables, cuyo registro aparece con el código respectivo para cada categoría observada de la covariable para el individuo (En la tabla 3.3 se especifica la codificación de cada ua de las categorías de las covariables).

En este mismo sentido, para el cliente con $id = 2$, se registra su información en las filas número 2, 3, 4 y 5 . Este cliente registra 3 recurrencias: la primera a los 852 días de instalado, la segunda, a los 1153 dias y la tercera a los 1302 días desinstalándose completamente a los 1659 días. Nótese que las columnas *fecha_inicio*y *fecha_fin* registran períodos de tiempo consecutivos. La primera fecha de la columna *fecha_inicio* para el cliente con $id = 2$, es la fecha de ingreso de este cliente a la empresa de telecomunicaciones (el 12/04/2006). La *fecha_fin* de esta misma fila, muestra la fecha en que el cliente con $id = 2$ tuvo su primera desinstalación de uno de sus servicios (el 11/08/2008). En la fila siguiente, se tiene para estas dos mismas variables y para el mismo cliente, en *fecha_inicio* la fecha de desinstalación del primer servicio (11/08/2008) y en *fecha_fin* la fecha de desinstalación de otro de sus servicios (el 08/06/2009) y así hasta registrar los cuatro períodos de tiempo que determinan los tiempos inter-ocurrencias para el cliente en mención. Al finalizar el tiempo de estudio, éste era un cliente desinstalado totalmente por lo que se registran tres indicadores de ocurrencia de desintlaciones (1 en la variable *evento*) y una indicadora de ocurrencia en la variable *terminal*en el ultimo tiempo de desconexión que sucede el 27/10/2010. Los tiempos inter-ocurrencias registrados para este cliente son: 852, 301 y 149. El último tiempo registrado (357), corresponde al tiempo entre la última recurrencia y la ocurrencia del evento terminal. Este cliente, era ciudad: CALI (Código 2); segmento: ESTRATÉGICAS (Código 1); grupo objetivo: NO ES DE UN HOLDING (Código 2); Número de servicios: MÁS DE 2 (Código 2); Número de quejas: [1-20](entre 1 y 20, Código 2).

En el mismo sentido de lectura de las tablas, el cliente 3 registra una desinstalación a los 1932 días de ingresado a la empresa y desinstalación total a los 1999 días de instalado. El único tiempo inter-ocurrencias registrado para este cliente es: 1932. El último tiempo registrado (67), corresponde al tiempo entre la última recurrencia y la ocurrencia

TABLA 3.3. Covariables usadas en el modelamiento.

COVARIABLE	Descripción	Categorías (entre paréntesis código de la categoría)
SEGMENTO	Segmento de mercado del cliente	Estratégica(1), Grande(2), Intermedia(3)
CIUDAD	Ciudad de la sede principal del cliente	Bogotá(1), Cali(2), Medellín(3), Barranquilla(4), Bucaramanga(5), Otras(6)
NRO_SERV	Nro de servicios con los que cuenta el cliente	Menos o igual a 2 (1), Más de 2 (2)
NRO_QUEJAS	Nro de quejas por servicio puestas por el cliente	0 (1), $[1-20]$(2), $[21-100]$(3), >101(4)
$GRUP\_OBJ(Grupoobjetivo)$	Indicadora de la pertenencia de la empresa a un holding de empresas	Es de un holding (1), No es de un holding(2)

del evento terminal. Este cliente, era ciudad: BARRANQUILLA(Código 3); segmento: INTERMEDIAS (Código 3); grupo objetivo: NO ES DE UN HOLDING (Código 2); Número de servicios: MENOS O IGUAL A 2 (Código 1); Número de quejas: [21-100](entre 21 y 100, Código 3).

El cliente 4 registra 3 desinstalaciones y al final del tiempo de estudio aún no se ha desintalado completamente, por lo que es un individuo censurado para el estudio ($terminal = 0$).

El cliente 5 permaneció en la compañía 1995 días y fué desinstalado completamente el día de finalización del estudio.

Un esquema que representa estos datos se muestra en la figura 3.3.

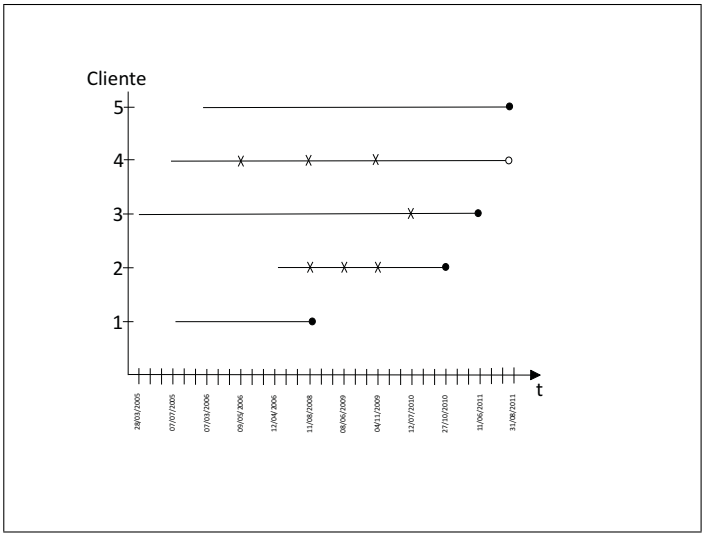

FIGURA 3.3. Gráfica muestra datos estudio.

3.1.7. El modelamiento

Las covariables en consideración para el modelamiento fueron dispuestas de tipo ordinal, la descripción de cada una de ellas, al igual que la codificación usada para el programa, se muestra en la tabla 3.3.

Todas las variables fueron codificadas de manera que se tuviera el sentido ordinal de las mismas. Esto se dispuso de esta manera, ya que, en general, para el manejo de este segmento empresarial, por ser tan especializado, se dedican esfuerzos que siempre son medibles desde el punto de vista de inversión presupuestal: para la variable SEGMENTO, los clientes estratégicos son los que más valen (para cualquier estrategia) ya que son quienes hacen la mayor inversión en telecomunicaciones. En este segmento se encuentran principalmente el sector financiero y de las telecomunicaciones. En este mismo sentido siguen los clientes grandes y los intermedios.

La variable ciudad se toma como ordinal por dos razones: 1) La participación de mercado que tiene la compañía en estas ciudades (Que es más alta según el orden en que aparecen listadas las ciudades en la tabla 3.2), y 2) La inversión presupuestal que se hace en planes de comunicación y de ampliación de fuerza de ventas, dada también por la penetración obtenida según el orden mencionado en el numeral anterior. Se asume como accionable porque como se menciona, dependiendo de la ciudad las acciones de comunicaciones y de inversión en campañas de mercadeo son diferentes, en el caso de encontrar que una ciudad tiene un impacto importante en el riesgo, éstas estrategias pueden ser modificadas.

El número de servicios y el número de quejas, se codifican y se transforman a variables de tipo ordinal. La pertenencia a un grupo objetivo se toma como ordinal, por que el hecho de no pertenecer a un holding de empresas le dá menor valor al cliente desde el punto de vista de atención comercial e interna de la compañía: un cliente que pertenece a un holding tiene mucho mejor servicio que uno que no pertenece.

La información del histórico de recurrencias por cliente se obtuvo para un total de 2457 clientes. La tasa de clientes perdidos (muerte) es del 25 %. El número promedio de recurrencias por cliente es de 2,31. La cantidad de clientes que no tuvieron ninguna desinstalación en el período de estudio son 1840 .

La programación para correr los modelos se hizo siguiendo la propuesta dada por Lu (2008) . El programa utilizado se encuentra en el Apéndice 1. Además, se tuvieron en cuenta las siguientes consideraciones para los datos:

- Se descartaron los clientes que fueron desconectados por problemas de no pago o por procesos con cartera, por considerarse que estos clientes no hacían parte del problema de desconexión que genera la pérdida del ingreso del cliente cuando éste decide dejar la empresa por causa voluntaria para tomar los servicios con otro proveedor. Los clientes desconectados por problemas con cartera o por no pago, se consideran pérdidas de cliente involuntaria.

- Los clientes que tenían más de 100 desconexiones en el tiempo de estudio no se incluyeron en el análisis pues éstos son clientes que normalmente trasladan servicios de un lugar a otro, luego, estas desconexiones no se toman como pérdidas de ingreso.

- Los tiempos inter-ocurrencias entre el evento terminal y la última desconexión de servicio -antes del evento terminal - que eran menores a 30 días se excluyeron y solo se tuvo en cuenta al tiempo hasta el evento terminal. Esto se debe a que una desconexión definitiva de un cliente corporativo puede demorar hasta 30 días, es decir, los tiempos excluidos comúnmente son tiempos de desconexiones definitivas del cliente.

TABLA 3.4. Coeficientes del modelo de riesgo para las recurrencias - con todas las variables.

COVARIABLE	Parámetro	Estimador	Error estándar	GL	Valor t	$Pr > \|t\|$
$SEGMENTO$	β_1	-0.2812	0.03495	2451	-8.05	< 0.0001
$CIUDAD$	β_2	0.05381	0.02175	2451	2.47	0.0134
NRO_SERV	β_3	-0.01940	0.03144	2451	0.62	0.5372
NRO_QUEJAS	β_5	-0.00728	0.01024	2451	-0.71	0.4773
$GRUPO\_OBJ$	β_4	-0.05733	0.03927	2451	1.46	0.1444

TABLA 3.5. Coeficientes del modelo de riesgo para el evento terminal- con todas las variables.

COVARIABLE	Parámetro	Estimador	Error estándar	GL	Valor t	$Pr > \|t\|$
$SEGMENTO$	α_1	0.3450	0.07849	2451	4.40	< 0.0001
$CIUDAD$	α_2	0.08565	0.03125	2451	2.74	0.0062
NRO_SERV	α_3	-0.2884	0.1118	2451	-2.58	0.0099
NRO_QUEJAS	α_5	-0.1015	0.02335	2451	-4.35	< 0.0001
$GRUPO\_OBJ$	α_4	0.2646	0.1495	2451	1.77	0.0768

El modelamiento conjunto se conforma por los dos modelos:

$$\lambda_i(t) = \nu_i \lambda_0(t) \exp(\beta_1 * SEGMENTO + \beta_2 * CIUDAD + \beta_3 * NRO\_SERV + \beta_4 * GRUPO\_OBJ + \beta_5 * NRO\_QUEJAS) \quad (3.3)$$

$$\gamma_i(t) = \nu_i^{\eta} \gamma_0(t) \exp(\alpha_1 * SEGMENTO + \alpha_2 * CIUDAD + \alpha_3 * NRO\_SERV + \alpha_4 * GRUPO\_OBJ + \alpha_5 * NRO\_QUEJAS) \quad (3.4)$$

Al estimar el modelo conjunto, según la teoría dispuesta en el capítulo 2, los coeficientes asociados con el proceso de las desinstalaciones, es decir, el modelo para el proceso de **eventos recurrentes** son los que se listan en la tabla 3.3; y los coeficientes del modelo de riesgo de pérdida de cliente, es decir, los asociados al proceso del **evento terminal** obtenidos mediante el modelamiento propuesto, se listan en la tabla 3.4.

La estimación de η, el coeficiente que determina la relevancia de las recurrencias sobre el evento terminal y de θ que es la estimación de la varianza para el término de fragilidad, se presentan en la tabla 3.5.

La tabla 3.5 muestra que los coeficientes de las variables NRO DE SERVICIOS, NRO DE QUEJAS Y GRUPO OBJETIVO, no son significativos (al 5 %) para el modelo de

TABLA 3.6. Estimaciones para el término de fragilidad - modelo con todas las variables.

Parámetro	Estimador	Error estándar	GL	Valor t	$Pr > \|t\|$
η	1.1603	0.3574	2451	3.25	0.0012
θ	0.3210	0.02225	2452	14.43	< 0.0001

Tabla 3.7. Coeficientes finales del modelo de riesgo para las recurrencias.

COVARIABLE	Parámetro	Estimador	Error estándar	GL	Valor t	$Pr > \|t\|$	HR
SEGMENTO	β_1	-0.2578	0.03143	2451	-8.20	< 0.0001	0.7727

Tabla 3.8. Coeficientes finales del modelo de riesgo para el evento terminal.

COVARIABLE	parámetro	Estimador	Error estándar	GL	Valor t	$Pr > \|t\|$	HR
SEGMENTO	α_1	0.3752	0.06761	2451	5.55	< 0.0001	1.45
NRO_SERV	α_3	-1.3083	0.1225	2451	-10.68	< 0.0001	0.27
NRO_QUEJAS	α_5	-0.1870	0.04401	2451	-4.25	< 0.0001	0.83

las recurrencias (p=0.53,0.47 y 0.54 respectivamente), mientras que para el del evento terminal solamente deja de ser significativa la variables GRUPO OBJETIVO ($p = 0.07$).

Puesto que se encuentran variables no significativas en los modelos ajustados, éstas se excluyen del modelamiento. El coeficiente de la variable CIUDAD, para el modelo de las recurrencias en el intento de modelamiento excluyendo las variables no significativas resulta ser no significativo por lo que también esta variable es excluída. De esta manera se obtienen los estimadores finales de los coeficientes para los dos modelos, los cuales se muestran en las tablas 3.6 y 3.7.

Las estimaciones de η y de θ se presentan en la tabla 3.8.

Dado que η es significativo al 5 %, se puede afirmar que el proceso de las recurrencias de desinstalaciones es informativo del evento terminal, lo que indica que hay un nivel asociación entre estos dos procesos y que el hecho de que los clientes tengan desconexiones recurrentes impacta significativamente (y positivamente $\eta = 1.0946$) el riesgo de pérdida de los clientes. Ahora, θ es también altamente significativo, es decir, la varianza de los ηs (término de fragilidad compartido) no es estadísticamente igual a cero (los ηs no son idénticamente 1), resultado que también informa del hecho de que la heterogeneidad observada en los procesos asociados a los eventos tanto recurrentes como terminal no se debe únicamente a las covariables incluídas en el modelamiento. $\theta = 0.2329$ indica que la tasa de desinstalaciones no es muy variable entre los clientes.

La única variable significativa para el proceso de recurrencias es la variable SEGMENTO. A medida que el segmento de la empresa es menos alto, la tasa de desintalaciones decrece cerca de un 23 % ($HR = 0.7727$). Es un resultado esperado, toda vez que los clientes que más enlaces/servicios poseen, son los clientes más grandes (segmento estratégicas o grandes). El segmento también tiene un impacto relevante en la tasa de riesgo al evento terminal, a medida que el segmento es menos alto, la tasa de riesgo de desintalación aumenta en un 45 % ($HR = 1.45$) aproximadamente. Por ejemplo los clientes intermedios que son los clientes de menor valor (segmento más bajo) en el conjunto de clientes incluídos en el estudio, se desinstalan más fácilmente que los de los segmentos más grandes

Tabla 3.9. Estimaciones final para el término de fragilidad.

Parámetro	Estimador	Error estándar	GL	Valor t	$Pr > \|t\|$
η	1.0946	0.2410	2451	4.54	< 0.0001
θ	0.2329	0.02418	2451	9.63	< 0.0001

TABLA 3.10. Estadísticos de ajuste para los dos modelos

Estadístico de ajuste	Modelo con todas las variables	Modelo con las variables significativas
Verosimilitud $-2log$	55.533	55.529
AIC	55.597	55.583
AICC	55.597	55.583
BIC	55.783	55.740

por contar con menos servicios. Normalmente también es menos compleja la migración de sus servicios a otro operador.

Las variables número de servicios y número de quejas también son significativas para el modelo de riesgo del evento terminal. En cuanto los clientes tienen más servicios, la tasa de desinstalación total del cliente decrece en un 73 % ($HR = 0.27$) -la adquisición de más servicios genera niveles de fidelidad en los clientes - y a medida que el número de quejas crece, la tasa de desintalación total decrece aproximadamente en un 17 % ($HR = 0.83$). Esta última interpretación puede ser objeto de discusiones, pero puede ser corroborada ya que muchos clientes corporativos recurren a la posibilidad de poner quejas constantemente para generar alarmas de atención de tal manera que sean solucionados sus casos rápidamente. Sin embargo por otros análisis se sabe que en el intervalo de 6 a 10 quejas hay un nivel de deserción no despreciable. Esta información será objeto de más estudio para un próximo modelamiento.

La exclusión de las variables no significativas aumenta la significancia del término η conservándose la interpretación de la validez del modelamiento conjunto de los dos procesos.

Los estadísticos de ajuste para los dos modelos se mantienen muy similares (tabla 3.9). Para el modelo con las variables significativas, todos los indicadores resultan ser un poco más pequeños, demostrando este resultado un mejor ajuste del modelo. Además, todos se encuentran dentro del mismo rango de valor, garantizando la convergencia del modelo final.

3.1.8. Evaluación de la idoneidad del modelo

El modelo ajustado está enmarcado teóricamente dentro del contexto del modelo de riesgos proporcionales de Cox, por seguir el modelo AG. En este sentido es necesario evaluar esta hipóteis. Para esto, se sigue la sugerencia del método gráfico propuesto en Andersen et al. (1982) , donde se sugiere dibujar las curvas del riesgo estimado acumulado ($\hat{\Lambda}_s(t)$) vs t, para las covariables del modelo con diferentes niveles de estratificación. El gráfico debe presentar lineas rectas provenientes del mismo origen.

En los gráficos 3.4 y 3.5 se muestra el resultado del riesgo estimado acumulado para las recurrencias ($\hat{R}(t)$) y el riesgo estimado acumulado para el evento terminal ($\hat{\Lambda}(t)$) vs t, para la variable SEGMENTO.

El gráfico 3.4 muestra la tendencia del riesgo para los tres segmentos ESTRATÉGICAS, GRANDES e INTERMEDIAS . Son claras las diferencias de las líneas de los riesgos estimados para las recurrencias, sugiriendo el cumplimiento de supuesto de proporcionalidad para la variable SEGMENTO. No se observan las mismas tendencias en esta variable para los riesgos del evento terminal graficados en la figura 3.5, ya que se sabe que para

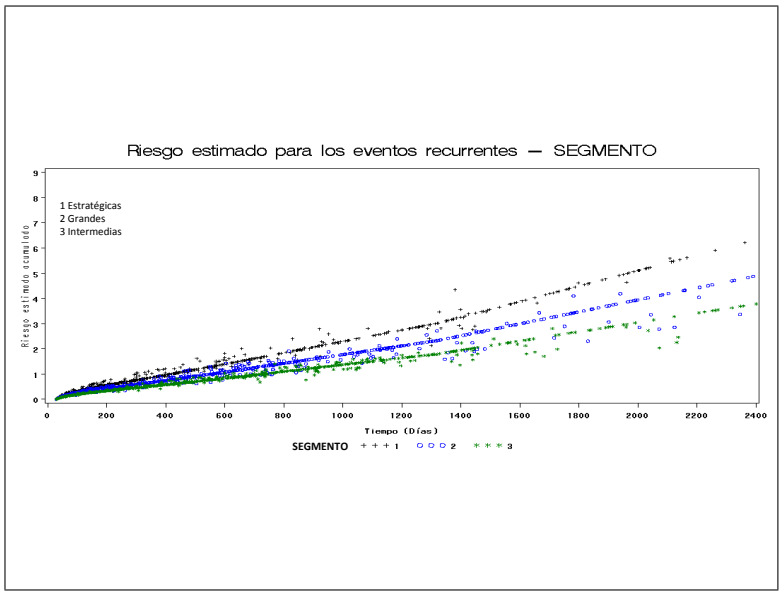

FIGURA 3.4. Función de riesgo para los eventos recurrentes - variable SEGMENTO

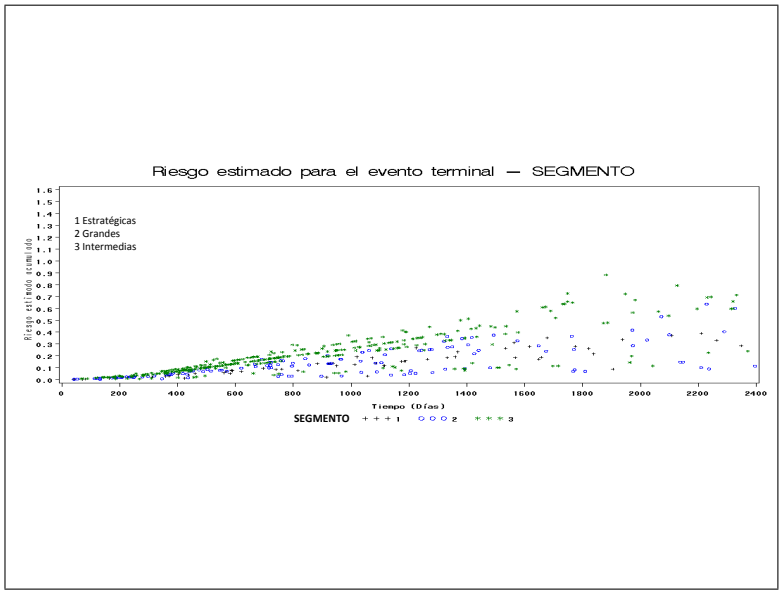

FIGURA 3.5. Función de riesgo para el evento terminal - variable SEGMENTO

este modelo los efectos alestorios causan algunas fluctuaciones del riesgo es decir, puntos más lejanos en el tiempo pueden obtener riesgos más pequeños que puntos más cercanos al punto cero en el tiempo de estudio (Lu & Liu (2008)). Sin embargo la variable segmento no se elimina del modelo por su nivel de significancia en el mismo.

En las figuras 3.6 y 3.7 se presentan los mismos gráficos descritos en el párrafo anterior, pero esta vez para la variable NÚMERO DE QUEJAS. Para observar la proporcionalidad esperada para diferentes estratificaciones de la variable, se utilizó la misma estratificación con la que se hizo el modelamiento, percibiéndose en el gráfico de $\hat{\Lambda}(t)$ vs t (gráfico 3.6),

lineas rectas desde el origen para los estratos 1 (0 quejas), 2 (entre 1 y 20 quejas) y 3 (entre 21 y 100 quejas) con algunos riesgos dispersos, mientras que el estrato 4 (más de 101 quejas) no presenta la misma tendencia (figura 3.5). En los primeros modelos ajustados se particionó esta variable en 8 niveles que mostraban mayor desagregación del NRO DE QUEJAS. Dado que en las pruebas de idoneidad para evaluar este mismo aspecto de proporcionalidad con estos 8 estratos se observaba una dispersión más grande de los riesgos estimados, se optó por realizar varias agregacones en donde ésta última es la que muestra mejor ajuste como se observa en el gráfico analizado , y al menos para tres de 4 los estratos de la variable.

En el gráfico 3.7 ($\hat{R}(t)$ vs t), se observa el no cumplimiento del supuesto de proporcionalidad para la variable NRO DE QUEJAS. La proporcionalidad en este modelo la domina la variable SEGMENTO de tal manera que la estratificación del número de quejas no se hace visible y queda prácticamente superpuesta sobre las líneas de riesgo de la variable mencionada. Los riesgos graficados para la variable NRO DE QUEJAS quedan traslapados sobre las tres líneas de riesgo de la variable SEGMENTO. Esto corrobora la no inclusión de la variable NRO DE QUEJAS en el modelo de riesgo para las recurrencias.

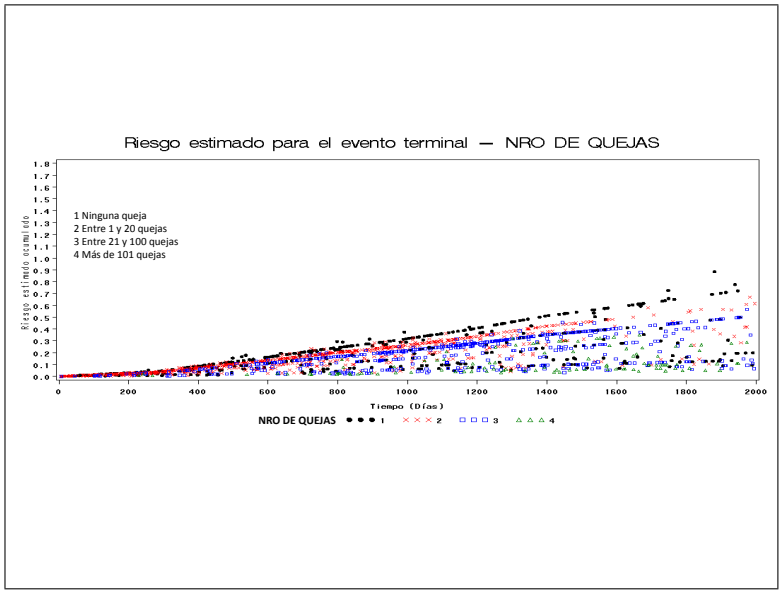

FIGURA 3.6. Función de riesgo para el evento terminal - variable NRO DE QUEJAS

Finalmente, en las gráficas 3.8 y 3.9 se dibujan igualmente los riesgos estimados acumulados para los dos procesos, el de recurrencias y el terminal respectivamente, para la variable NÚMERO DE SERVICIOS. Para la construcción de los gráficos se utilizan de la misma manera que las figuras anteriores las estratificaciones usadas para el modelamiento. La figura 3.9 tiene la misma intepretación que la figura 3.7, por lo que similarmente, este resultado corrobora la exclusión de esta variable en el modelo de eventos recurrentes. En la gráfica 3.8 se muestran varias líneas rectas de riesgos para el evento terminal, que indicarían las tendencias de riesgo de una estratificación subyacente, pero no se observa la forma esperada del gráfico con la estratificación usada para la variable en cuestión (los dos estratos usados son los que representan las 2 categorías: 2 o menos servicios y 3 o más servicios). Sin embargo, el ajuste del modelo y la significancia de la variable en el mismo, proponen que esta variable se tenga en cuenta para el modelo.

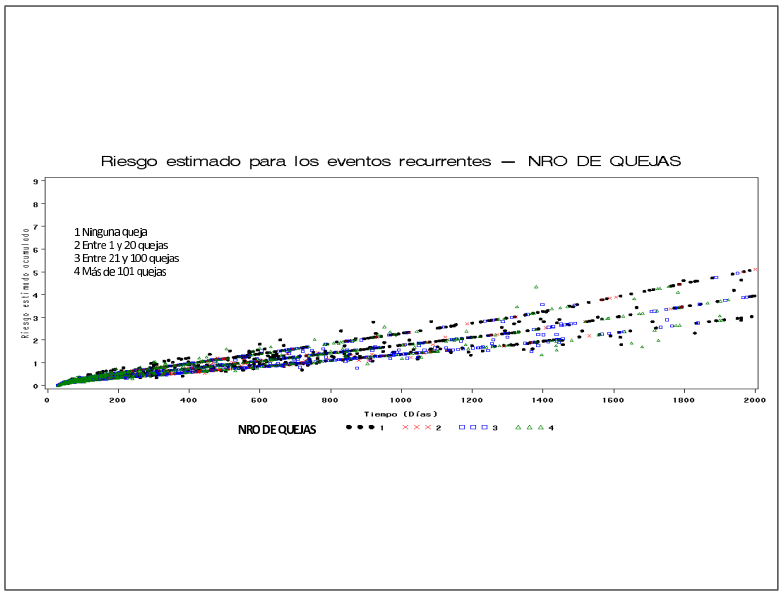

FIGURA 3.7. Función de riesgo para los eventos recurrentes - variable NRO DE QUEJAS

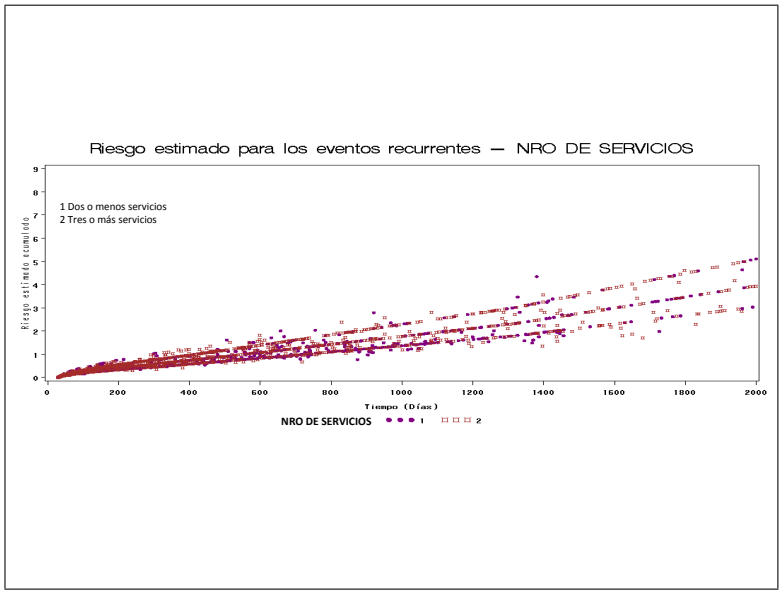

FIGURA 3.8. Función de riesgo para los eventos recurrentes - variable NRO DE SERVICIOS

Para la validación de la distribución del término de fragilidad estimado ($\Gamma(1,\theta)$) se sabe que dado que no se puede hacer la separabilidad del efecto de la función del riesgo base en la estimación del término de fragilidad, se hace complejo evaluar este supuesto (Liu et al. (2004)), sin embargo se cumple que la media es igual a 1.

Otro mecanismo de validación del modelo consistió en la evaluación de las tasas de-serción reales un año después de ajustado el modelo para los 1840 clientes de la base que fueron censurados por no haber sufrido el evento terminal en el tiempo bajo estudio. De esto se obtuvo una tasa de deserción global del 8 % (147 clientes habían desertado de la

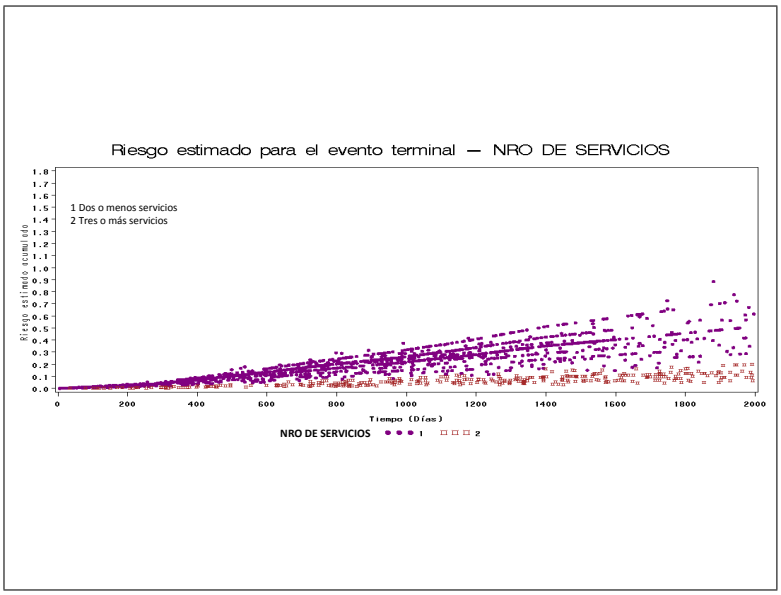

FIGURA 3.9. Función de riesgo para el evento terminal - variable NRO DE SERVICIOS

empresa al año de haber ajustado el modelo). Por las variables incluídas en el modelamiento, para la variable SEGMENTO se encontró una tasa de deserción del 12.2 %, 25.17 % y 62.6 % para los clientes ESTRATÉGICAS, GRANDES e INTERMEDIAS, respectivamente corroborando lo anunciado en el modelo de mayor riesgo de deserción a medida que decrece el segmento (mayor riesgo para las intermedias, menor para las estratégicas); para la variable NRO DE QUEJAS, la incidencia de deserción fue de 20.41 % para los clientes que no registraron quejas, 51.02 % para los que interpusieron entre 1 y 20 quejas, 24.49 % para los que interpusieron entre 21 y 100 quejas y 4.08 %, para los que registraron más de 100 quejas. este resultado también es concordante con el análisis realizado acerca de que a mayor número de quejas el riesgo de deserción decrece.

Finamente para la variable NRO DE SERVICIOS, se obtuvieron las tasas de deserción de 86.39 % y 13.61 % para las categorías clientes con 2 o menos servicios y clientes con 3 o más servicios respectivamente, que igualmente concuerda con lo analizado para los coeficientes del modelo donde se indicaba que la tasa de deserción instantánea para el conjunto de clientes que tuvieran 2 o menos servicios es de manera importante, mucho más alta que los que tuvieran instalados 3 o más servicios.

El modelo conjunto así evaluado, cumple las expectativas del modelamiento y se considera un modelo apropiado para el objetivo de gestión de clientes desde las dos perspectivas: deserción pasiva y deserción total.

CAPÍTULO 4

Conclusiones y recomendaciones

Conclusiones

El modelo conjunto ajustado en este trabajo, determinó un nivel de dependencia del proceso de desintalaciones de un cliente con el proceso de pérdida total del mismo. Éste nivel de dependencia sugiere empezar a hacer un tratamiento de gestión de retención o fidelizaciòn de clientes desde que estos entran en el proceso de desconexión repetitiva de sus enlaces o servicios instalados.

En cuanto a las variables significativas obtenidas en el ajuste de los dos modelos tanto para el proceso de deserción pasiva (la desinstalación repetitiva de los servicios) como de pérdida total del cliente (la desinstlación total del cliente) se obtuvo que la variable SEGMENTO es significativa para los dos procesos. Siendo imperativo hacer notar que la estimación del coeficiente β de esta variable para el proceso de recurrencias, indica que a medida que el cliente es de menor valor, la tasa de desinstalaciones es menor, mientras que para el proceso de ocurrencia del evento terminal pasa lo contrario: a medida que el cliente es de menor valor, la tasa de ocurrencia de riesgo de pérdida total de cliente es mayor. Esto es, es más probable perder un cliente del segmento INTERMEDIAS (Cliente de menor valor) que un cliente del segmento ESTRATÉGICAS (cliente de mayor valor) y es menos probable observar desinstalaciones de un cliente del segmento INTERMEDIAS, que de un cliente del segmento ESTRATÉGICAS.

En resumen, la tasa de ocurrencia de pérdida total de cliente es más alta a medida que el cliente es de menor valor y la tasa de ocurrencia de desinstalaciones de servicios es más alta a medida que el cliente es de mayor valor. Este resultado sugiere un evaluación en términos de pérdida de ingreso, ya que se sabe que un cliente de mayor valor puede desinstalar servicios, lo cuales podrían tener mayor valor que uno o varios clientes del segmento de menor valor. Puesto que se sabe por experiencia que un cliente de mayor valor puede durar mucho tiempo desinstalándose completamente por la complejidad de sus servicios, en términos de controlar la fuga de ingreso, según este resultado, la sugerencia es que se determine un proceso de gestión para este tipo de clientes con alto riesgo de deserción pasiva.

En el modelo ajustado para el proceso de ocurrencia del riesgo terminal, dos variables más toman relevancia en determinar el riesgo de pérdida total de clientes: el número de

servicios del cliente y el número de quejas interpuesto en el período de estudio. El efecto estimado de la variable número de servicios es negativo, indicando esto que en la medida que el cliente tiene mayor número de servicios menor es la tasa de pérdida total del cliente. En cuanto al impacto del número de quejas en la tasa de desinstlación total de clientes, se observa que disminuye en la medida que el cliente interpone mas quejas. Esta lectura corrobora la sospecha de que un número alto de quejas interpuesto por un cliente no es determinante de que es más alta su probabilidad de deserción. Se sugiere hacer un análisis de esta variable que determine puntos de inflexión con respecto a la probabilidad de riesgo de deserción total del cliente.

Finalmente, en cuanto a la bondad del modelo, el ajuste del modelo propuesto para el caso de modelamiento de los dos procesos de interés, mejora la expectativa de modelamiento que se puede tener mediante el ajuste del modelo de Cox, en el sentido de poder evaluar el nivel de dependencia de los dos procesos y verificar la hipótesis de significancia del parametro ν que se ajustó en el modelamiento para tal necesidad de información. Además este tipo de modelamiento también da respuesta al interés de encontrar un nivel de heterogeneidad en el modelo para el proceso de riesgo de pérdida del cliente que permite plantear la hipótesis de considerar más covariables para el ajuste de este modelo.

Además con el modelo conjunto ajustado, al igual que con un ajuste del modelo de Cox, se cubrió también la expectativa de encontrar un modelo que diera cuenta de los dos procesos en términos de la evaluación de la significancia de covariables incluídas para el modelamiento, situación que permite poder hacer un modelo de gestión de clientes más apropiado teniendo en cuenta las covariables de más impacto para cada modelo ajustado.

Recomendaciones

En el recorrido analítico para determinar el modelo que finalmente aquí se propone, se consideró ajustar un modelo de riesgos en competencia el cual no se adoptó ya que normalmente en los sistemas de información no es recabada la causa de retiro de los clientes y sin esta información es imposible poder ajustar este tipo de modelo.

Para la estimación de los parámetros se usó en la función de verosimilitud los tiempos inter-ocurrencias, ya que se observó un mejor ajuste de los datos usando éstos en lugar de los tiempos calendario. Se sugiere que sea ajustado el modelo mediante las dos vías para validar con qué tipo de tiempos (inter-ocurrencias o calendario) ajusta mejor el modelo.

En un trabajo futuro se recomienda considerar la inclusión de más variables ya que en este modelamiento, por no tener la posibilidad de llegar a esta información del cliente, no fue posible incluir por ejemplo, un factor que puede estar altamente correlacionado con los dos procesos estudiados que es la satisfacción del cliente con la compañía.

APÉNDICE A

Anexo 1

A.1. Macro en SAS para modelar eventos recurrentes con evento terminal

Se presenta la macro en SAS usada para el modelmiento de los datos. Aquí:

- *inds*es el archivo de datos. Aquí cada línea es una ocurrencia (recurrencia) del evento de interés. Éste archivo debe incluir una variable que indica si el registro es una recurrencia, censura o evento terminal.
- *idvar* Es la variable identificador del individuo.
- *timevar* Es la variable tiempo.
- *statusvar* Es la variable del status del evento (0 si es censura, 1 si es evento recurrente y 2 si es evento terminal.
- *covar* Son las covariables.
- *inpar* Es el archivo que contiene los valores iniciales de los parámetros
- *parest* Es el archivo de salida que contiene los parámetros estimados.
- *nu_est* Es el archivo de salida que contiende los efectos aleatorios estimados para cada individuo.
- *cumh* Es el archivo de salida que contiene las estimaciones de la función de riesgo acumulado de eventos recurrentes.
- *outS* Es el archivo de salida que contiene las estimaciones de la función de sobrevida del evento terminal.

La macro finalmente se corre con el comando:

%recurr(Inds, Idvar, timevar, statusvar, covar, inpar, parest, nu_est, cumh, outS);

```
/* PROGRAMA DEFINITIVO:  1 VARIABLE PARA EL MODELO DEL RIESGO A LAS RECURRENCIAS Y  3 VARIABLES
                         PARA EL MODELO DEL RIESGO TERMINAL*/

%macro recurr(inds,idvar,timevar,statusvar,covar,covar1,covar2,inpar,parest,nu_est,cumh,outS);

/* Obteniendo cuantiles para la construcción de la función de riesgo base para los eventos recurrentes*/;

proc univariate data=&inds(where=(&statusvar=1)) noprint;
 var &timevar;
 output out=quant_r pctlpts=0 10 20 30 40 50 60 70 80 90 100 pctlpre=qr;
run;


/* Obteniendo cuantiles para la  construcción de la función de riesgo base para el evento muerte */;

proc univariate data=&inds (where=(&statusvar=0 or &statusvar=2)) noprint;
     var &timevar;
     output out=quant_d pctlpts=0 10 20 30 40 50 60 70 80 90 100 pctlpre=qd;
run;

proc transpose data=quant_r out=quant_r2;
run;
data _null_;
     length a $ 150;
     retain a ' ';
     set quant_r2 end=last;
     a= trim(a)||' '||col1;
     if last then call symput('quant_r',a);
run;

proc transpose data=quant_d out=quant_d2;
run;
data _null_;
     length a $ 150;
     retain a ' ';
     set quant_d2 end=last;
     a= trim(a)||' '||col1;
     if last then call symput('quant_d',a);
run;

/* Calcular la duración de cada intervalo de cuantil, para el indicador de evento en cada intervalo */;

data all;
     set &inds;
     array quant_r {11} _TEMPORARY_ ( &quant_r);
     array quant_d {11} _TEMPORARY_ ( &quant_d);

     array dur_r {10} dur_r1-dur_r10;
     array dur_d {10} dur_d1-dur_d10;

     array even_r {10} even_r1-even_r10;
     array even_d {10} even_d1-even_d10;
```

```
      do i=1 to 10;
          dur_r{i}=0;
          dur_d{i}=0;
          even_r{i}=0;
          even_d{i}=0;
      end;

/* Para el evento recurrente */;
      if event=1 then do;
           do i=2 to 11;
                if &timevar<=quant_r{i} then do;
                     even_r{i-1}=1;
                     dur_r{i-1}=&timevar-quant_r{i-1};
                     i=11;
                end;
                else dur_r{i-1}=quant_r{i}-quant_r{i-1};
           end;
      end;

      else do; /* Si es muerto o censurado */

           do i=2 to 11;
                if &timevar<=quant_d{i} then do;
                     even_d{i-1}=(event=2);
                     dur_d{i-1}=&timevar-quant_d{i-1};
                     i=11;
                end;
                else dur_d{i-1}=quant_d{i}-quant_d{i-1};
           end;
      end;
run;
ods output ParameterEstimates=&parest;

/* Comando del PROC con los parámetros de corrida del PROC NLMIXED*/
proc nlmixed data=all qpoints=30 maxiter=800 maxfunc=5000 noad;
 parms / data=&inpar; /*inpar es el archivo de parámetros de entrada. Aquí se define  los valores iniciales de los
coefiientes beta(terminal), alpha(recurrentes) , gamma y vara*/

bounds r01 r02 r03 r04 r05 r06 r07 r08 r09 r10 h01 h02 h03 h04 h05 h06 h07 h08 h09 h10 vara >=0;

/* Cálculo del riesgo base y riesgo base acumulado para los eventos  recurrentes */
base_haz_r=r01*even_r1+r02*even_r2+r03*even_r3+r04* even_r4 +
   r05*even_r5+r06*even_r6+r07*even_r7+r08*even_r8
   +r09 * even_r9 + r10 * even_r10;
     cum_base_haz_r=r01*dur_r1+r02*dur_r2+r03*dur_r3+r04*dur_r4 +
       r05*dur_r5+r06*dur_r6+r07*dur_r7+r08*dur_r8+r09*dur_r9 +
       r10 * dur_r10;

 /* Cálculo del  riesgo base  y riesgo base acumulado para el evento muerte */
      base_haz_d=h01*even_d1+h02*even_d2+h03*even_d3+h04*even_d4 +
       h05*even_d5+h06*even_d6+h07*even_d7+h08*even_d8+
       h09 * even_d9 + h10 * even_d10;
      cum_base_haz_d=h01 * dur_d1 + h02 * dur_d2 + h03 * dur_d3 +  h04 * dur_d4 + h05 * dur_d5 +
        h06 *dur_d6 + h07 * dur_d7 +  h08* dur_d8 +h09 * dur_d9 + h10 * dur_d10;
```

```
/* Transformación para que los términos  aleatorios  tengan la distribución Gamma(1,Theta)*/
     p=cdf('NORMAL', nu);
     if p > 0.999999 then p= 0.999999;
     g2=quantile('GAMMA',p,1/vara);
     g=g2*vara;  /* g se distribuye gamma con media 1 y varianza vara */

    /* Definición de  los modelos*/

    mu1= beta1 * &covar + log(g); /* para los eventos recurrentes */

    mu2= alpha1 * &covar + alpha3 * &covar1 + alpha5 * &covar2 + gamma * log(g);  /* Para el evento muerte*/

    loglik1=-exp(mu1) * cum_base_haz_r;
    loglik2=-exp(mu2) * cum_base_haz_d;

     /*log verosimilitud para el evento recurrente */

     if event=1 then loglik=log(base_haz_r) + mu1+loglik1 +loglik2 ;

    /*log verosimilitud para el evento muerte(terminal) */

     if event=2 then loglik=loglik1 +log(base_haz_d)+mu2+loglik2;

    /*log verosimilitud para la censura */

     if event=0 then loglik=loglik1 + loglik2;

    /*Comando para el modelo*/
     model &timevar ~ general(loglik);

    /*Definición de la distribución del término de fragilidad */
     random nu  ~ normal(0,1) subject=&idvar out=&nu_est;

    /*Predicciones  (tasa de riesgo ) para los eventos recurrentes y terminal*/

     predict -loglik2 out=&outs; /*Riesgo estimado acumulado para el evento muerte*/
     predict -loglik1 out=&cumh;   /*Riesgo estimado acumulado para el evento recurrente*/
run;
%mend;
```

APÉNDICE B

Anexo 2

B.1. Transformación para obtener que la distribución de los efectos aleatorios en el PROC NLMIXED sea $\Gamma(1, \theta)$

En este apéndice, se sigue la propuesta de Nelson et al. (2006).

Supóngase que los efectos aleatorios que se quieren incluir en el modelo (que se asumen contínuos) tienen una distribución no normal $f(b_i, \theta)$ y que necesariamente el programa sobre el cual se quiere hacer el ajuste del restringe a que la distribución de los éstos sea normal (como sucede con el PROC NLMIXED). Sea a_i un efecto aleatorio proveniente de una distribución normal estándar, esto es $a_i normal(0, 1)$. Entonces, usando la transformación integral de probabilidad, $u_i = \Phi(a_i)$ tiene una distribución uniforme (0,1), donde $\Phi(.)$ es la función de distribución normal estándar acumulada (FDA). Aplicando la transformación integral de probabilidad una vez más $F_\theta(b_i)$ también tiene una distribución uniforme (0,1) donde $F_\theta(.)$ es la función de distribución normal estándar acumulada de b_i, con parámetro θ. Se sigue entonces $b_i = F_\theta^{-1}(u_i)$ tiene densidad $f(b_i, \theta)$, donde $F_\theta^{-1}(.)$ es la FDA inversa de b_i. Entonces $b_i = F_\theta^{-1}(\Phi(a_i))$ tiene la función de distribución no-normal de interés.

Condicionado al efecto aleatorio (o fragilidad) b_i, se asume: 1) b_i es normalcon media 0 y varianza θ, y 2) $b_i = log(g_i)$, donde $g_i > 0$ tiene la distribución gama:

$$f(g_i \mid \theta_1, \theta_2) = g_i^{1/\theta_1 - 1} \exp(-g_i/\theta_2)[\Gamma(1/\theta_1)\theta_2^{1/\theta_1}] \tag{B.1}$$

Por identificabilidad se toma $\theta_2 = \theta_1$, así que g_i tiene media 1. Esto es:

$$E(g_i) = \theta_2/\theta_1 = 1 \tag{B.2}$$

Así, B.1 se reduce a:

$$f(g_i \mid \theta_1) = g_i^{1/\theta_1 - 1} \exp(-g_1/\theta_1)/[\Gamma(1/\theta_1)\theta_1^{1/\theta_1}] \tag{B.3}$$

De esta manera, para lograr los efectos aleatorios con distribución gama cuando se ajusta el modelo mediante el PROC NLMIXED de SAS es necesario hacer el siguiente conjunto de tranformaciones:

1. $a_i \sim N(0,1)$
2. $p_i = \Phi(a_i)$
3. Hágase $g_{i2} = F_{\theta_1}^{-1}(p_i)$
4. $g_i = \theta_1 g_{i2}$

El PROC NLMIXED usa $\theta_2 = 1$, pero en el programa se pueden definir los parámetros θ_1 y θ_2.

Como interpretación para el caso de múltiples observaciones por individuo (cluster), cuando θ_1 se acerca a 0, las observaciones dentro de un cluster son independientes, mientras valores grandes de θ_1 indican correlaciones altas dentro del cluster (individuo).

Bibliografía

[1] D.A. Aaker and A.L. Biel, *Brand equity & advertising: Advertising's role in building strong brands*, Lawrence Erlbaum Associates, 1993.

[2] O. Aalen, O. Borgan, and H.K. Gjessing, *Survival and event history analysis*, Springer, 2008.

[3] O.O Aalen, *Nonparametric inference for a family of counting processes*, Ann. Statist. **6** (1978), 701–726.

[4] L.J.S.M. Alberts, *Churn prediction in the mobile telecommunications industry - an application of survival analysis in data mining*, Master's thesis, Maastricht University, 2006.

[5] P.K. Andersen, O. Borgan, R.D. Gill, and N.Keiding, *Statistical models based on counting processes*, Springer, 1993.

[6] P.K. Andersen and R.D. Gill, *Cox's regression model for counting processes: a largue sample study*, Annals of Statistics **10** (1982), 1100–1120.

[7] P.K. Andersen and N. Keiding, *Multi-state models for event history analisys*, Statistical Methods in Medical Research **11** (2002), 91–115.

[8] H.P. Blossfeld, K. Golsh, and G. Rohner, *Event history analysis with stata*, Lawrence Erlbaum Associates, Inc., 2007.

[9] M. Braun and D. Schweidel, *Modeling customer lifetimes with multiple causes of churn*, Tech. report, Massachusetts Institute of Technology y University of Wisconsin - Madison, 2011.

[10] J. Castañeda and B. Gerritse, *Appraisal of several methods to model time to multiple events per subject: Modelling time to hospitalizations and death*, Revista Colombiana de estadística **33** (2010), no. 1, 43–61.

[11] R. Cook and Lawless, *The statistical analysis of recurrent events*, 2006.

[12] R.D. Cox, *Regression models and life tables (with discussion)*, Journal on the Royal Statistical Society. Series B (1972), 187–220.

[13] F.G. Dolivo, *Counting processes and integrated conditional rates: a martingale approach with application to detection theory*, Ph.D. thesis, University od Michigan, 1974.

[14] R.D. Gill, *Censoring and sthocastic integrals*, Mathematical Centre Tracts 124. Mathematisch Centrum Amsterdam (1980).

[15] S. Hung and D. Yen, *Applying data mining to telecom churn management*, Expert Sistems with Applications **31** (06), no. 3, 515–524.

[16] O.B. Kalhida, B. Sunarti, A. H. Norazina, and B. Faizin, *Data mining in churn analysis model for telecommunications industry*, Journal of Statistical Modeling and Analytics **1** (2010), 19–27.

[17] E.L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association **53** (1958), 457–481.

[18] J.P Klein and M.L. Moeschberger, *Survival analysis: Techniques for censored and truncated data*, Springer-Verlag, 1997.

[19] D. Kleinbaum and M. Klein, *Survival analysis a self-learning text*, 2 ed., 2005.

[20] G. Kraljevic and S. Gotovac, *Modelling data mining applications for prediction of prepaid churn in telecommunications services*, Automatika **3** (2010), 275–283.

[21] S. Lipsitz, N. Laird, and D. Harrington, *Using the jacknife to estimate the variancde of regression estimators from repeated measures studies*, Communication in Statistics. Theory and Methods **19** (1990), no. 1, 821–845.

[22] L. Liu, R. Wolfe, and X. Huang, *Shared frailty models for recurrent events and a terminal event*, Biometrics **60** (2004), 747–756.

[23] T. Louis, *Finding the observed information matrix when using the em algorithm*, Journal of the Royal Statistical Society (1982).

[24] J. Lu, *Predicting customer churn in the telecommunications industry - an application of survival analysis modeling using sas*, SUGI 27 (2002), 114–127.

[25] ———, *Modeling customer lifetime value using survival analysis - an application in the telecommunications industry*, SUGI 28 (2003).

[26] L. Lu and C. Liu, *Analysis of correlated recurrent and terminal events data in sas*, Statistic & Analysis NESUG 2008 (2008).

[27] T. Mutanen, *Customer churn analysis*, Research Report (2006).

[28] K.P. Nelson, S.R. Lipsitz, and otros, *Use of the probability integral transformation to fit nonlinear mixed-effects models with nonormal random effects*, American Statistical Association (2006).

[29] R.L. Prentice, B.J. Williams, and A.V. Peterson, *On the regression analysis of multivariate failure time data*, Biometrika **68** (1981), 373–379.

[30] G. Rodríguez, *Multivariate survival models*, Tech. report, Princeton, 2005.

[31] V. Rondeau, S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran, *Join frailty models for recurring events and death using maximun penalized likelihood estimation: aplication on cancer events*, Biostatistics **8** (2007), 708–721.

[32] S. Rosset, E. Neuman, U. Eick, and N. Vatnic, *Customer lifetime value models for decision support*, Data Mining and Knowledge Discovery, (2002).

[33] P.J. Smith, *Analysis of failure and survival data*, CHAPMAN & HALL, 2002.

[34] A. Tamaddoni, *Predicting customer churn in telecommunications service providers*, Master's thesis, Luleá University of Technology, 2009.

[35] T.M. Therneau and P.M. Grambsch, *Modeling survival data: Extending the cox model*, Springer, 2000.

[36] L.J. Wei, D.Y. Lin, and L. Weissfeld, *Regression analysis, of multivariate incomplete failure time data by modeling marginal distributions*, Journal of de American Statistical Association **84** (1989), 1065–1073.