PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

MODELO DE SUPERVIVENCIA DE LARGA DURACIÓN CON RIESGOS PROPORCIONALES Y ESTIMACIÓN DEL RIESGO BASE VÍA SPLINES: MODELAMIENTO DE ABANDONO DE SEGUROS

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Hector Mattos Galarza

Asesor: Victor Giancarlo Sal y Rosas Celi

Miembros del jurado:
Dr. Victor Giancarlo Sal y Rosas Celi
Dr. Cristian Luis Bayes Rodríguez
Dr. Luis Hilmar Valdivieso Serrano

Lima, Julio 2020

Dedicatoria

A mi alma mater, la Pontificia Universidad Católica del Perú, que guió mis pasos y el de muchos profesionales para aportar con nuestros conocimientos a la sociedad.

Agradecimientos

A mis padres, que nunca dejan de alentarme en cada reto que me propongo. A mi hermana por sus ganas de ayudar y a mi sobrino que me inspira a ser mejor cada día como su ejemplo. A mi asesor Giancarlo Sal y Rosas, que con su guía he podido culminar esta etapa. A mi gran amigo Jonatán Rojas, por su incondicional apoyo. Y a todos mis profesores de la maestría que hicieron muy gratificante esta fase.

Resumen

Palabras-clave: Supervivencia, Cox, Proporcional, Riesgo, Fracción de Cura, Splines.

Los modelos de supervivencia, aquellos que tratan de describir el tiempo a la ocurrencia de uno o más eventos, han demostrado tener gran versatilidad para poder modelar distintos tipos de eventos y un alcance mayor al que inicialmente se propuso. Su aplicación varía desde el área de la medicina hasta usos en actividades financieras como análisis de riesgos de activos, entre otros. Este trabajo tiene como motivación el análisis del tiempo de permanencia de un cliente con contrato de póliza de seguros. En esta aplicación, solo una fracción de los clientes son susceptibles a la terminación del contrato y, en este sentido, se requiere que el modelo cuente con la flexibilidad de asumir que no todos los clientes son susceptibles al evento de interés. En este trabajo, se propone un modelo de larga duración asumiendo un modelo de riesgos proporcionales para los clientes susceptibles de abandono y donde la función de riesgo basal de este último se modela vía funciones de splines monótonas. Este trabajo empieza con la definición del modelo, el proceso de estimación de parámetros, escenarios de simulación donde se evalúa el desempeño del proceso de estimación e inferencia y finalmente una aplicación para estudiar los factores asociados con el abandono de clientes en una compañía de seguros en el Perú.

Abstract

Keyword: Survival, Cox, Proportional, Hazard, Cure Fraction, Splines.

Survival models, those that are focused on trying to describe the time before the ocurrence of one or more events, have demonstrated great versatility in their capacity to model various types of events and a further reach than initially proposed. Its application encompasses from medical trials to uses in financial activities like assets risk management, among others. This work focuses in the analysis of the time of a customer until their decision of termination of an insurance policy. In this application, only a fraction of the population are prone to terminate their contract and, in this sense, it is needed that the model have a certain degree of flexibility of assuming that not all the clients are susceptible to this event. A long-term proportional hazard model is proposed in this work with base risk function modeled via monotone splines. This work starts with the model definition, the parameters estimation process, simulation scenarios where the estimation and inference process performance is evaluated and finally an application to study the associated factors with the churn process for an insurance company in Perú.

Índice general

Lista de Abreviaturas

f.d.	Función de densidad.
f.d.a.	Función de distribución acumulada.
v.a.	Variable aleatoria.
EMV	Estimador de máxima verosilimitud.
M.R.P.	Modelo de Riesgos Proporcionales.
EM	Algoritmo de Esperanza-Maximización.

Lista de Símbolos

β	Logaritmo del cociente de riesgo asociado a una covariable.
η	Coeficiente de la fracción de cura del modelo logístico.
γ_i	Coeficiente de la componente i de los splines.
δ	Indicador de censura.
$\mathcal{L}$	Función de verosimilitud.
$\lambda_0(t)$	Riesgo instantáneo base en el tiempo t.
$\Lambda_0(t)$	Riesgo acumulado base en el tiempo t.
T	Tiempo hasta la ocurrencia del evento de interés.
f	Función de densidad.
F, G	Función de distribución acumulada.
S	Función de supervivencia.
λ	Función de riesgo instantáneo.
δ	Indicador de censura.
$\mathbb{I}$	Función indicatriz.

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

El abandono de clientes en el rubro de seguros es un problema al que se enfrentan todas las empresas aseguradoras (Gerber et al., 2018). Esto afecta a los resultados de la empresa en varios aspectos como la disminución de las reservas para inversión, la recuperación de los gastos de adquisición, entre otros. Es así que una de las mayores dificultades que se presenta en este rubro es modelar de manera precisa el proceso de abandono de un cliente para poder prever, aprovisionar y retener a tiempo a los clientes.

Se plantea aplicar un modelo sobre el proceso de abandono, donde se asumirá que existen clientes que, por distintas razones dadas sus características, no son susceptibles a dicho proceso. Es así que en esta investigación se propone realizar un modelo de fracción de cura, donde se asume un modelo de riesgos proporcionales en las personas susceptibles a abandono y donde la función de riesgo basal se modela a partir de splines monótonas.

1.1. Antecedentes y Motivación

Las aseguradoras tradicionales basan varios de sus procesos financieros y comerciales en estimaciones de ciertas características de sus clientes. Entre estos procesos se observa el fijado de precios (Wang, 1995), el cual toma en cuenta el riesgo de toda la cartera de clientes a siniestrarse, o las acciones comerciales que intentan apuntar al sector del mercado que es más propenso a adquirir los servicios de la aseguradora (Harrison and Ansell, 2002). Es así que, debido a la necesidad de hallar estas características se incurre en la implementación de diversos análisis estadísticos.

Dado que es mas costoso atraer nuevos clientes que retenerlos, uno de los análisis más frecuentes es estudiar los factores asociados con la persistencia, o de forma complementaria, el abandono de la cartera (Morik and Köpcke, 2004). Este se realiza con el objetivo de fijar metas y comisiones, pronosticar demanda de los servicios, establecer el comprador objetivo, realizar campañas de retención de clientes, entre otros. Este análisis se realiza tradicionalmente a toda la cartera de clientes o segmentado por alguna característica muy general como el rango etario o el canal de venta, lo que restringe la obtención de conclusiones específicas sobre cada individuo y del impacto de sus características sobre su comportamiento de abandono del seguro (Chiang, 1984).

En la actualidad, para este tipo de análisis se utilizan regresiones binomiales con una ventana de tiempo fija para la ocurrencia del evento que se quiere analizar (Sabbeh, 2018). Sin embargo, este enfoque limita bastante la información que se pueda adquirir del modelo. Es así que si, por ejemplo, se quiere saber cuanto es el tiempo esperado hasta que el evento ocurra, se deben hacer algunas suposiciones, como la independencia del evento con respecto al tiempo, que no necesariamente son válidas y podrían influir los resultados.

Motivado por estas restricciones, en esta investigación se propone el uso de modelos de supervivencia para el análisis del tiempo hasta el abandono de seguros. En la actualidad se tienen varios ejemplos de este tipo de aplicación que incluyen la aplicación de modelos paramétricos, semi-paramétricos y no paramétricos (Larivière and Van den Poel, 2004; Wong, 2011). Dada la gran cantidad de datos disponible en la actualidad y a la necesidad de obtener información sobre cómo se comporta el abandono de los individuos de acuerdo a sus características, se propone un modelo semi-paramétrico que permita el uso de covariables, ya que brinda la flexibilidad suficiente para capturar el comportamiento sin perder la interpretabilidad de sus resultados.

El modelo propuesto es un modelo de regresión sobre el tiempo hasta la ocurrencia del evento de ínteres bajo la presencia de datos censurados por la derecha y donde no todos los individuos son propensos a desarrollar el evento de ínteres. Este método de regresión, además de permitir el cálculo de la probabilidad de abandono en una ventana de tiempo como se hace actualmente, también permite hablar del tiempo esperado hasta el evento e incluso saber en que instante se tiene mayor propensión o riesgo a abandonar el seguro.

El modelo específico, para las personas susceptibles a desarrollar el evento de interés, será un modelo de riesgos proporcionales, donde la función de riesgo acumulado basal será descrita por una familia de funciones splines monótonas. Esto permite que el riesgo acumulado basal sea modelado con precisión sin importar su naturaleza y además brinda un estimador más suave y fácil de utilizar, ya que en general, la aplicación de este tipo de regresiones debe ser sencilla en su uso e implementación para su adopción en la empresa. Por otro lado, la parte proporcional de la función de riesgo del modelo y la probabilidad de pertenencia a la fracción de la población que no está expuesta al evento son paramétricos, ya que es muy importante para el negocio conocer qué caracteriza al comportamiento de sus clientes y cómo puede accionar estos factores importantes para tomar mejores decisiones.

Más aún, dada la naturaleza de los seguros, existe la posibilidad de que el evento de abandono no ocurra. Es por esto que se propone analizar la incidencia del evento como una componente del modelo a la cual se llamará fracción de cura. Esta componente no observada extrae el efecto de si el evento puede ocurrir o no.

Este tipo de modelos son conocidos como modelos de supervivencia con fracción de cura o de larga duración. A partir de la investigación de los diversos modelos existentes, se puede ver una gran variedad de desarrollos como el uso de un modelo paramétrico log-gamma generalizado con fracción de cura el cual utiliza una regresión log-gamma generalizada modificada para permitir la posibilidad de supervivencia a larga duración (Ortega et al., 2009). Este modelo abarca, como casos especiales, los modelos de regresión log-exponencial, log-Weibull y log-normal. Otros tipos de modelo que se han desarrollado recientemente tienen en consideración diversos tipos de procesos por los cuales la información puede estar limitada o restringida, lo que hace que algunos eventos no sean observables (Shen et al., 2019; Calsavara et al., 2019). En el caso a desarrollar, solo se tiene información restringida donde el evento de interés sucede luego del fin de experimento, por lo que no se indagará más en el desarrollo de este tipo de modelos. Asimismo, han habido avances en el desarrollo de modelos de supervivencia con fracción de cura utilizando splines que permiten una estimación de riesgo base más suave sobre el tiempo (Bremhorst and Lambert, 2016).

Por otro lado, para la aplicación del modelo en investigación se tendrán ciertas consideraciones previas que nos ayudarán en el desarrollo de la misma. La primera es que solo se observará a cada

póliza del seguro hasta su primer abandono. Esto quiere decir que, si esta póliza es rehabilitada, su información solo se tomará hasta la primera vez que caducó. Otro punto importante a considerar son las pólizas grupales o contratadas por empresas. Estas tienen una estructura distinta y, por ende, el proceso por el cual se cancelan difiere del que se quiere modelar. Es por esto que solo se considerarán las pólizas que han sido contratadas por personas naturales.

Por otro lado, es importante tener en cuenta la independencia de las observaciones. Este punto es fácil de aceptar dado que no es nada extraño que las razones de abandono sean independientes entre sí. Asimismo, asumimos que los factores que definen si una persona puede incurrir en abandono o no, es decir, las covariables que entran a la componente de cura del modelo, son independientes del tiempo observado.

1.2. Objetivos

El objetivo general de la tesis es el estudio de un modelo con fracción de cura asumiendo un modelo de riesgos proporcionales para su población susceptible y donde la función de riesgo basal, de este último, se modela vía splines monótonas. De manera específica:

- Revisar el estado del arte en el análisis de supervivencia con fracción de cura
- Proponer, estudiar e implementar el modelo sujeto de investigación
- Revisar el comportamiento del modelo planteado usando escenarios simulados
- Aplicar el modelo a un conjunto de datos de clientes de una empresa de seguros
- Comparar el modelo propuesto con otros modelos conocidos

Capítulo 2

Conceptos Preliminares

En el presente capítulo se tocarán los conceptos necesarios para el desarrollo de esta investigación.

2.1. Tiempo de Supervivencia

Sea T la variable aleatoria que describe el tiempo a la ocurrencia de un evento de interés. Esta se asume tiene una función de densidad (f.d) y función de distribución acumulada (f.d.a) dadas por f y F, respectivamente. La f.d.a está definida por

$$F(t) = \int_0^t f(s)ds. \tag{2.1}$$

Por otro lado, se define la función de supervivencia como el complemento de la función de distribución acumulada. De manera más específica

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t). \tag{2.2}$$

Un concepto importante en el entendimiento de una variable aleatoria T es su función de riesgo instantánea. La función de riesgo instantánea, denotada por $\lambda(\cdot)$, está definida por

$$\lambda(t) = \lim_{h \to 0} \frac{P(T \in [t, t+h) \mid T \geq t)}{h}. \tag{2.3}$$

Finalmente, la función de riesgo acumulada está definida por la integral de la función de riesgo instantánea. De manera específica

$$\Lambda(t) = \int_0^t \lambda(s)ds. \tag{2.4}$$

Es importante notar algunas relaciones importantes entre estas tres funciones presentadas. En particular

$$f(t) = \frac{dF}{dt} \tag{2.5}$$

$$S(t) = e^{-\Lambda(t)} \tag{2.6}$$

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{2.7}$$

para $t > 0$.

2.2. Datos Censurados

Existen circunstancias donde el tiempo a la ocurrencia del evento no es medible pero sí es conocido si es menor o mayor que un tiempo observado. Este tipo de estructura de datos es conocida como datos censurados y es observable, ya sea por la estructura de diseño del estudio o por eventos que se consideran aleatorios.

Existen varios tipos de datos censurados. Los mas conocidos son: i) censura por derecha si el evento no se llega a observar; ii) censura por izquierda, si al primer control el evento ya sucedió y iii) censura por intervalo, si el evento de interés sucede entre controles. Los modelos de supervivencia permiten extraer la distribución del tiempo a la ocurrencia del evento de ínteres aun cuando algunas o gran parte de las observaciones esta sujetas a censuras. En la Figura 2.1 se puede observar la estructura de datos censurados para los tipos descritos anteriormente.

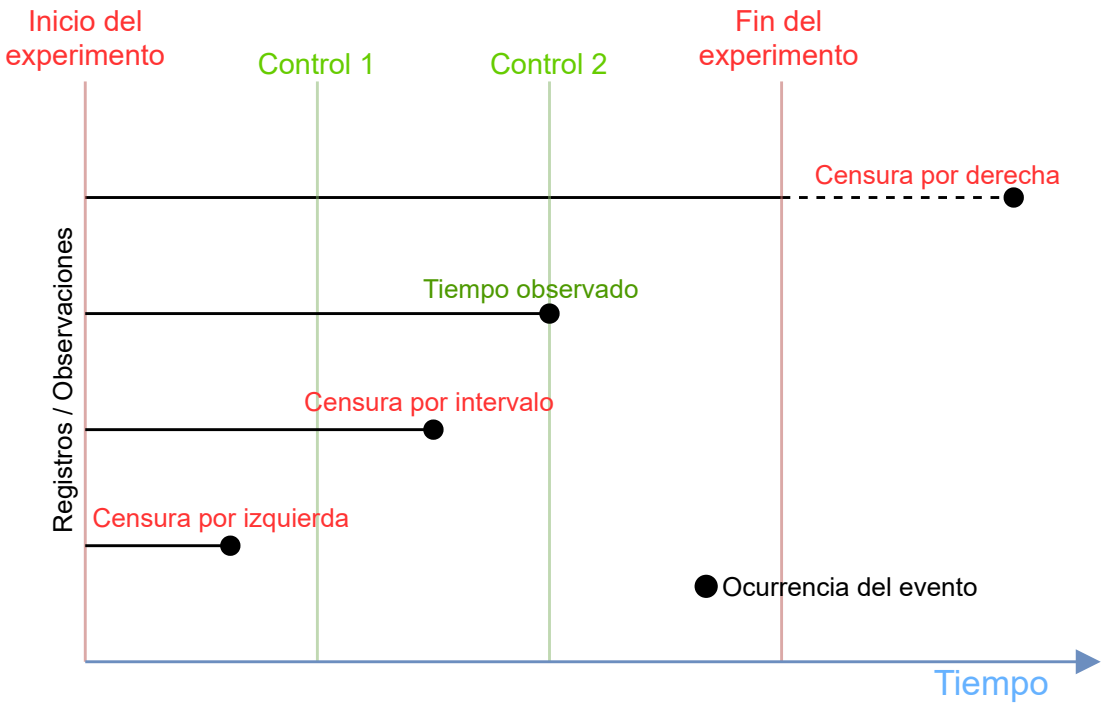


Figura 2.1: Estructura del tiempo observado

La presente investigación tratara en especifico el caso de datos censurados por derecha. En este contexto, sea $C \sim G$ el tiempo a que suceda la censura con f.d. and f.d.a. denotadas por g y G, respectivamente. Dado el tiempo a censura, se define el tiempo observado por $\bar{T} = min(C, T)$ y el indicador de censura por $\Delta = \mathbb{I}(T \leq C)$.

Dado los datos observados $\left\{\bar{T}, \Delta\right\}$, y asumiendo independencia entre T y C, se tiene la f.d. de $\bar{t}$ condicional a δ , denotada por w, por

$$w(\bar{t}|\delta = 1) = f(\bar{t})P(C \geq \bar{t}) = f(\bar{t})\left(1 - G(\bar{t})\right) \tag{2.8}$$

$$w(\bar{t}|\delta = 0) = g(\bar{t})P(T > \bar{t}) = g(\bar{t})\left(1 - F(\bar{t})\right) \tag{2.9}$$

A partir de (2.8) y (2.9) se tiene que la siguiente función de densidad conjunta para $\bar{T}$ y Δ:

$$w(\bar{t}, \delta) = \left[f(\bar{t})\left(1 - G(\bar{t})\right)\right]^{\delta} \left[g(\bar{t})\left(1 - F(\bar{t})\right)\right]^{1-\delta}. \tag{2.10}$$

2.3. Modelo de riesgos proporcionales

Una de las formas de caracterizar los modelos de supervivencia es a partir de la función de riesgo instantánea presentada en la Sección 2.1. Cox (1972) presentó lo que se conoce como el modelo de riesgos proporcionales el cual se puede escribir como

$$\lambda(t \mid X) = \lambda_0(t) \exp\left(X^{\mathsf{T}}\beta\right) \tag{2.11}$$

donde t representa el tiempo, $\lambda(t \mid X)$ es la función de riesgo instantáneo, $X = (X_1, \ldots, X_k)^{\mathsf{T}}$ es el vector de covariables, $\beta = (\beta_1, \ldots, \beta_k)$ es un vector de coeficientes que miden el impacto de cada covariable y $\lambda_0(t)$ es el riesgo instantáneo base o basal en el tiempo t.

Este modelo es de riesgos proporcionales ya que, como se puede observar, las covariables definen el riesgo como una proporción del riesgo base. En particular, β_i es el logaritmo del cociente de riesgos cuando la covariable X_i aumenta en una unidad. La Figura 2.2 muestra el supuesto del modelo de riesgos proporcionales para una covariable dicotómica.

Cabe resaltar la existencia de otros tipos de modelos de supervivencia, como el de tiempo de falla acelerados que no exhiben riesgos proporcionales (Kalbfleisch and Prentice, 2011).

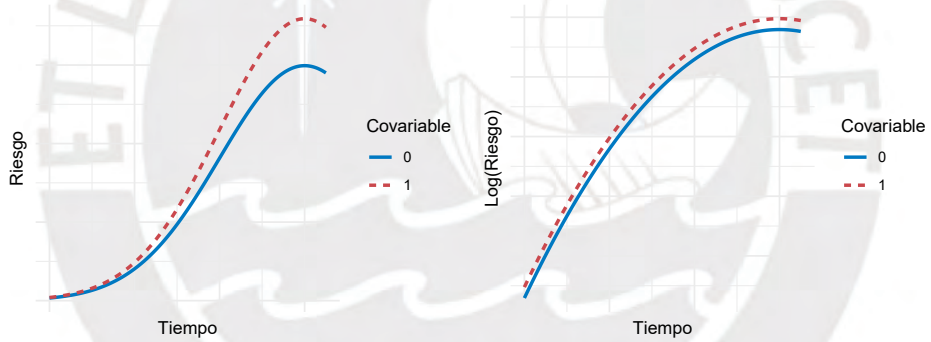


Figura 2.2: Modelo de riesgos proporcionales: Función de riesgo en escala natural (izquierda) y escala logarítmica(derecha)

2.4. Modelo de Mixtura y Fracción de Cura

Pierce et al. (1979) propusieron el uso de modelos de mixtura para modelar el comportamiento de poblaciones distintas. Esto puede ser físicamente significativo en algunos casos para su interpretación, ya que, según sus observaciones, en algunos experimentos existía una población que no parecía ser susceptible al evento de interés. Boag (1949), Berkson and Gage (1952) y Haybittle (1965) hicieron algunos avances iniciales sobre este tipo de modelos donde se establece el concepto de población curada y en Mould and Boag (1975) se puede ver una aplicación de la metodología en la búsqueda de tratamientos en pacientes con cáncer cervical. Farewell (1977) consideraron una generalización de estas técnicas, donde combina la relación logística para estudiar factores asociados a la probabilidad de ser suceptible al evento de ínteres y un modelo exponencial para estudiar la distribución del tiempo de incidencia.

Se procede a definir el problema de forma matemática. Se define la variable binaria Y, donde $Y = 0$ significa que el individuo no presentará el evento de interés, es decir, pertenece a la fracción de la población curada, y, por el contrario, $Y = 1$ representa que el individuo estará suceptible a que le suceda el evento de ínteres. Entonces, la probabilidad de que el evento ocurra en un tiempo mayor a t esta dada por

$$P(T > t) = P(T > t \mid Y = 1)P(Y = 1) + P(T > t \mid Y = 0)P(Y = 0). \tag{2.12}$$

De esta manera, el tiempo transcurrido hasta el evento para la población donde $Y = 1$ se puede modelar de distintas maneras, por lo que definimos $f(t|Y = 1, x)$ como la función de densidad del tiempo hasta el evento de dicha población. De la misma manera se define la probabilidad de que el tiempo T de ocurrencia del evento sea mayor que cualquier valor dado de t para la población donde $Y = 0$ es siempre 1 como se expresa en la siguiente ecuación

$$P(T > t|Y = 0, x) = 1\ , \forall t > 0 \tag{2.13}$$

Entonces, dada la proporción de la población que son susceptibles al evento de interés, denotado por $\pi = P(Y = 1)$, esta dada por

$$P(T > t) = \pi P(T > t \mid Y = 1) + 1 - \pi \tag{2.14}$$

La Figura 2.3 presenta una explicación gráfica de la idea de los modelos de larga duración. En estas se puede notar que se estudia el tiempo a la ocurrencia del evento en una sub población.

2.5. Splines Monótonas

Los polinomios de la forma $p(x) = \sum_{i=1}^{k} a_i x^{i-1}$ tienen gran importancia práctica en la estadística y matemáticas en general. Esto se debe en gran parte a dos características muy importantes: son lineales en el parámetro a_i a ser estimado y la combinación lineal de los $\{x^{i-1}\}$ son fáciles de manipular de manera algebraica y numérica, en especial con respecto a la derivación e integración. Sin embargo, los polinomios sufren de una gran limitante: la falta de flexibilidad para cambiar el comportamiento de la función alrededor de un valor dado x^* sin afectar el comportamiento de la función en cualquier otro valor x^{**}.

Los splines son funciones especiales definidas por partes bajo polinomios. En este sentido, los splines polinómicos cumplen con las características expresadas anteriormente y cuentan con mayor flexibilidad al construir la función $p(\cdot)$ a partir de polinomios unidos por partes.

Wegman and Wright (1983) presentaron múltiples usos de splines aplicados en la estadística como, por ejemplo, en el ámbito de regresiones no paramétricas. Ramsay et al. (1988) presentó la aplicacion de los M-splines, los cuales son una familia de mucho interés de splines polinómicos. Expresó que los splines polinómicos vienen dados por la combinación lineal $p(x) = \sum_{i=1}^{n} a_i M_i(x)$ donde n es el número de dimensiones del espacio funcional asociado al spline con grado y condiciones de continuidad definidos. Asimismo, $M_i(x)$ está definido dentro de un intervalo $[t_i, t_{i+1}]$, donde $t_{i+1} \geq t_i$. Los valores de t_i son conocidos como nodos y, más aún, todos los nodos distintos al máximo y mínimo nodo son denominados nodos internos. En particular, si se define un spline con el máximo nivel de continuidad $k - 1$, n toma el valor del número de nodos internos más el orden k del spline. En específico, para

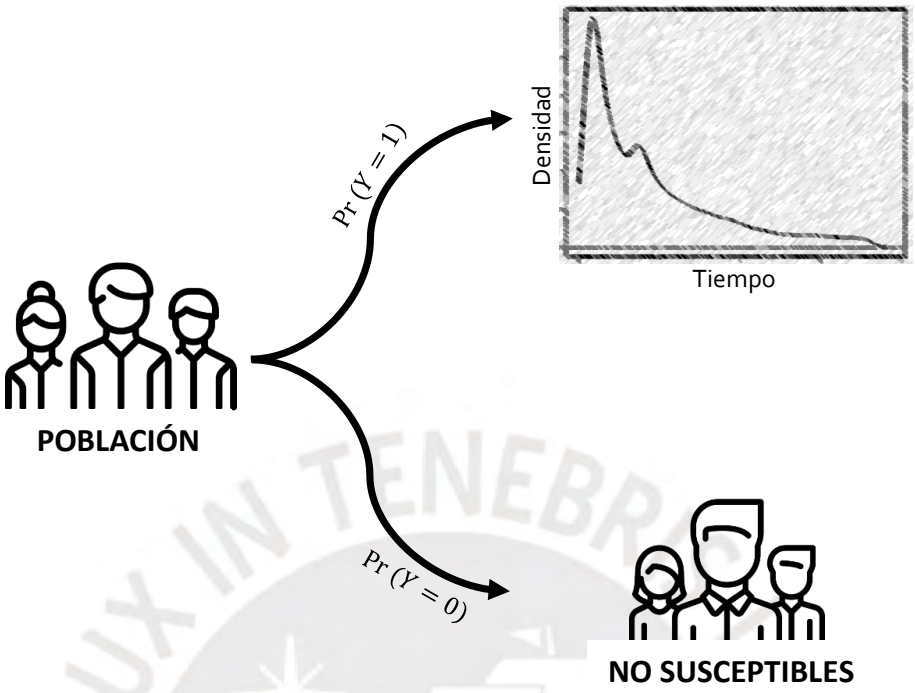


Figura 2.3: Modelo de mixtura

los M-splines, cada $M_i(x),\ i = 1, \ldots, n$ está definida como positiva para cualquier valor $t \in [t_i, t_{i+1}]$ y 0 para cualquier otro valor fuera del rango. Adicionalmente, $M_i(x)$ cuenta con la normalización $\int_{t_i}^{t_{i+1}} M_i(x)dx = 1$ (Curry and Schoenberg, 1965). No se entrará a más detalle en la definición matemática de los M-splines dado que no es el objetivo de este trabajo, pero se recomienda revisar los trabajos de Ramsay et al. (1988) y Curry and Schoenberg (1965).

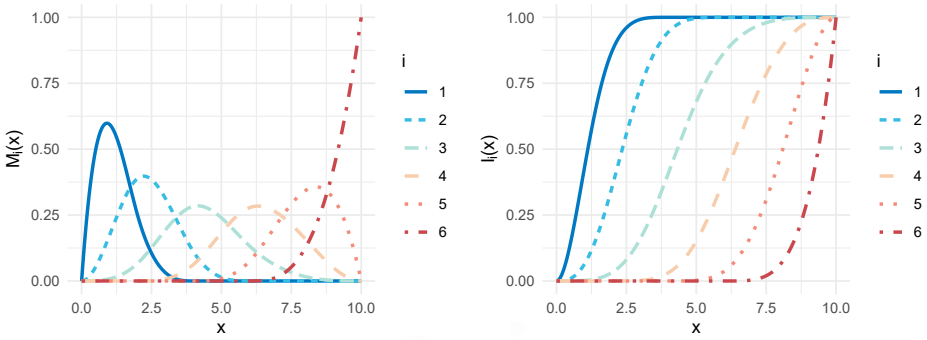
En función de esas referencias, es posible ver que todos los $M_i(x)$ en la familia de M-splines cuentan con las propiedades de funciones de densidad de probabilidad dentro del intervalo $[t_i, t_{i+1}]$. Más aún, dado que $M_i(x) = 0$ fuera del intervalo $[t_i, t_{i+1}]$ y positivo dentro de este, un cambio en el coeficiente a_i solo afectará los valores dentro del intervalo mencionado, lo que consigue la flexibilidad deseada de la sensibilidad al coeficiente de manera local.

Finalmente, se define la familia de I-Splines, denotados por $I_i(x)$ de grado d y k nodos están definidos como:

$$I_i(x) = \int_{t_i}^{x} M_i(u)du \tag{2.15}$$

Nótese que, por definición, los I-splines son funciones monótonas y que combinaciones lineales de estas, con coeficientes no negativos, son una base para la familia de funciones no decrecientes y no negativas. En la Figura 2.4a se puede observar las funciones descritas por las componentes de los M-Splines con 3 nodos en 2, 4 y 6 y de grado 3. Asimismo, en la Figura 2.4b se observan las componentes

de los I-Splines correspondientes. De esta manera es posible construir una función monótona creciente a partir de la combinación lineal de las componentes de una familia de I-Splines. Más adelante se verá cómo esto es útil para la construcción de una función de riesgo acumulado.



(a) Componentes de los M-Splines (b) Componentes de los I-Splines

Figura 2.4: Familia de Splines

Capítulo 3

Metodología

En este capítulo discutiremos el modelo a ser desarrollado. En particular, su construcción y proceso de estimación.

3.1. Modelo

Sea T el tiempo a la ocurrencia del evento de interés. Asimismo, se define una v.a. Y tal que $Y \sim Bernoulli(\pi)$, que caracteriza si una persona es propensa a desarrollar el evento de interés o no y donde π es la probabilidad de ser susceptible al evento de interés. Por nomenclatura se definirá

$$T = \begin{cases} T_S & , \quad Y = 1 \\ \infty & , \quad Y = 0 \end{cases} \tag{3.1}$$

donde T_S es el tiempo a desarrollar el evento para una persona susceptible. De esta manera, la función de supervivencia de dicha persona se puede escribir como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \pi P(T > t \mid Y = 1) + (1 - \pi) P(T > t \mid Y = 0) \\ &= \pi S^*(t) + 1 - \pi, \end{aligned} \tag{3.2}$$

donde $S^*(t)$ corresponde a la función de supervivencia de los susceptibles en el momento t. Más aún, el modelo propuesto asume proporcionalidad en la función de riesgo de las personas susceptibles (Cox, 1972). Este supuesto es expresado mediante la siguiente ecuación:

$$\Lambda^*(t|\boldsymbol{X}) = \Lambda_0^*(t) \exp(\boldsymbol{X}^\top \beta), \tag{3.3}$$

donde $\Lambda_0^*(t)$ es la función de riesgo basal acumulado en el tiempo t para los susceptibles. Sin embargo, este modelo considera una proporción de personas no susceptibles al abandono por lo que la función de riesgo acumulado para toda la población condicional a $\boldsymbol{X}$ y π queda reescrita como

$$\Lambda(t|\boldsymbol{X}, Y) = \begin{cases} \Lambda_0^*(t) \exp(\boldsymbol{X}^\top \beta), & \text{si } Y = 1 \\ 0, & \text{si } Y = 0 \end{cases}, \tag{3.4}$$

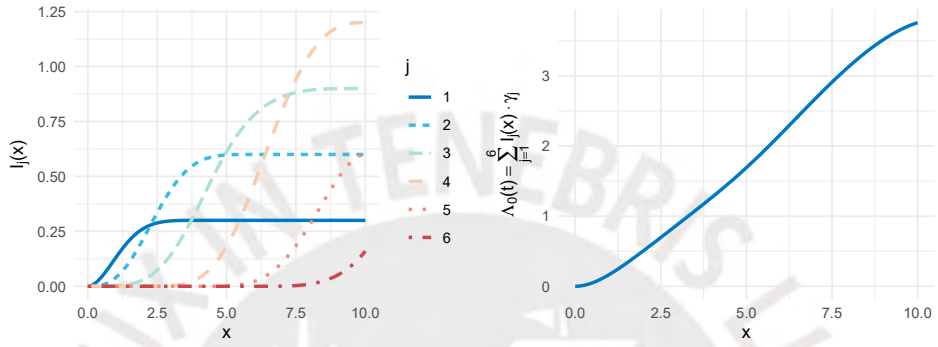
siendo $\Lambda_0^*(t)$ el riesgo basal acumulado en el tiempo t independiente de las covariables de la población susceptible a abandono.

En este trabajo se propone utilizarlas bases de splines monótonas definidas en el capítulo anterior (Ramsay et al., 1988).

De esta manera, se define el riesgo basal acumulado para la población susceptible como:

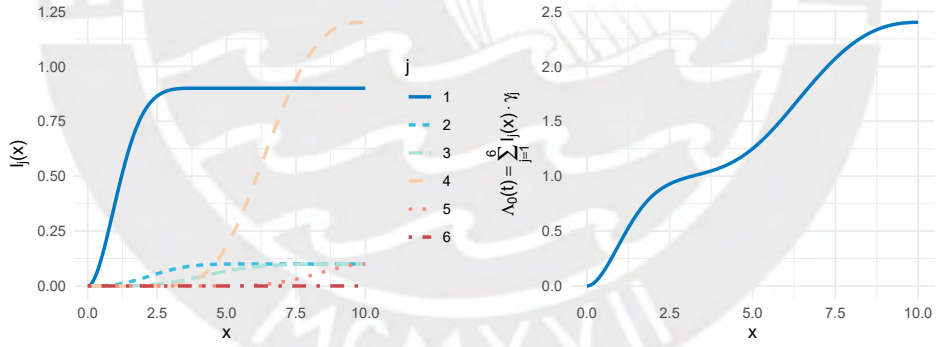
$$\Lambda_0^*(t) = \sum_{j=1}^{k} \gamma_j I_j(t|d,k), \tag{3.5}$$

donde dado que los `I-splines` son funciones no negativas, $\Lambda_0^*(t)$ es monótona siempre y cuando $\gamma_j \geq 0$, $\forall j \in \{1,\ldots,k\}$ (véase las Figuras 3.1 y 3.2 las cuales contienen ejemplos de los componentes de los I-splines).



(a) Componentes del I-spline (b) I-spline

Figura 3.1: Función de riesgo acumulado basal en función de I-splines con $d = 3$, $k = \{2,4,6\}$ y $\gamma_j = \{0{,}3, 0{,}6, 0{,}9, 1{,}2, 0{,}6, 0{,}15\}$ y un dominio $[0-10]$



(a) Componentes del I-spline (b) I-spline

Figura 3.2: Función de riesgo acumulado basal en función de I-splines con $d = 3$, $k = \{2,4,6\}$ y $\gamma_j = \{0{,}9, 0{,}1, 0{,}1, 1{,}2, 0{,}1, 0\}$ y un dominio $[0-10]$

Por otro lado, la probabilidad de ser suceptible al evento de ínteres será modelado en función de un conjunto de covariables $\boldsymbol{Z} = \left(Z_1, Z_2, \ldots, Z?k\right)$ vía un enlace logit. Es decir,

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{Z}^\top \eta \tag{3.6}$$

y su función de probabilidad estará dada por

$$P(Y = y|\boldsymbol{Z}) = \left(\frac{1}{1+\exp(-\boldsymbol{Z}^\top\eta)}\right)^y \left(\frac{\exp(-\boldsymbol{Z}^\top\eta)}{1+\exp(-\boldsymbol{Z}^\top\eta)}\right)^{1-y}, \tag{3.7}$$

donde η son los log-odds (parámetros desconocidos) que miden el impacto de las covariables $\boldsymbol{Z}$ sobre la probabilidad de pertenecer o no a la fracción no expuesta al evento.

Es así que, según (3.2) y (3.4), se tiene la siguiente ecuación para la función de supervivencia para toda la población.

$$S(t|\boldsymbol{X},\boldsymbol{Z}) = 1 + (\exp(-\Lambda^*(t|\boldsymbol{X})) - 1)P(Y=1|\boldsymbol{Z}) \tag{3.8}$$

Asimismo, a partir de la definición de la función de supervivencia $S(t|\boldsymbol{X},\boldsymbol{Z}) = 1 - F(t|\boldsymbol{X},\boldsymbol{Z})$ se define la f.d de T como

$$f(t|\boldsymbol{X},\boldsymbol{Z}) = \frac{-\partial S(t|\boldsymbol{X},\boldsymbol{Z})}{\partial t}. \tag{3.9}$$

A partir de (3.7), (3.8) y (3.9) y la relación $\frac{\partial I_j(t)}{\partial t} = M_j(t)$ se tiene la función de densidad del tiempo de supervivencia para toda la población

$$f(t|\boldsymbol{X},\boldsymbol{Z}) = \exp\left(-\sum_{j=1}^{k}\gamma_j I_j(t)\exp(\boldsymbol{X}^\top\boldsymbol{\beta})\right)\frac{1}{1+\exp(-\boldsymbol{Z}^\top\eta)}\left(\sum_{j=1}^{k}\gamma_j M_j(t)\exp(\boldsymbol{X}^\top\boldsymbol{\beta})\right). \tag{3.10}$$

En específico, para el caso especial sin covariables se obtiene la función de densidad

$$f(t) = \exp\left(-\sum_{j=1}^{k}\gamma_j I_j(t)\right)\frac{1}{1+\exp(-\eta_0)}\left(\sum_{j=1}^{k}\gamma_j M_j(t)\right). \tag{3.11}$$

Para hallar los estimadores de máxima verosimilitud, se necesita la expresión de la función de verosimilitud de toda la muestra. Dado que no toda la información es observada, se introduce el término δ_i que indica si el tiempo observado es de abandono $\delta_i = 1$ ó es una censura $\delta_i = 0$ para el individuo i. Es así que, dado los datos conocidos, se forma la siguiente función de verosimilitud:

$$\mathcal{L}(\theta|t_i,\delta_i,\boldsymbol{X}_i,\boldsymbol{Z}_i) = \prod_{i=1}^{n} f(t_i|\boldsymbol{X}_i,\boldsymbol{Z}_i)^{\delta_i} S(t_i|\boldsymbol{X}_i,\boldsymbol{Z}_i)^{1-\delta_i}, \tag{3.12}$$

con $\theta = \{\gamma,\beta,\eta\}$.

Para simplificar la expresión, se define $\psi_i = \exp(\boldsymbol{X}_i^\top\boldsymbol{\beta})$ y el conjunto de variables $\{T,\Delta,\boldsymbol{X},\boldsymbol{Z}\}$ como Γ. Es así que, reemplazando (3.8) y (3.10) en (3.12), la función de verosimilitud del modelo propuesto es dada por

$$\begin{aligned}\mathcal{L}(\theta|\Gamma) = \prod_{i=1}^{n}\Bigg[&\left(\exp\left(-\sum_{j=1}^{k}\gamma_j I_j(t)\psi_i\right)\frac{1}{1+\exp\left(-z_i^\top\eta\right)}\left(\sum_{j=1}^{k}\gamma_j M_j(t)\psi_i\right)\right)^{\delta_i}\\ &\left(1+\left(\exp\left(-\sum_{l=1}^{k}\gamma_j I_j(t_i)\psi_i\right)-1\right)\frac{1}{1+\exp\left(-z_i^\top\eta\right)}\right)^{1-\delta_i}\Bigg].\end{aligned} \tag{3.13}$$

Asimismo, la función de log-verosimilitud del modelo es dada por

$$\ell(\theta|\Gamma) = \sum_{i=1}^{n} \left[\delta_i \left(-\sum_{j=1}^{k} \gamma_j I_j(t_i)\psi_i \right) \right.$$
$$- \delta_i \log(1 + \exp\left(-z_i^\top \eta\right)) + \delta_i \log\left(\sum_{j=1}^{k} \gamma_j M_j(t_i)\psi_i \right)$$
$$\left. +(1-\delta_i) \log\left(1 + \frac{\exp\left(-\sum_{j=1}^{k} \gamma_j I_j(t_i)\psi_i\right) - 1}{1 + \exp\left(-z_i^\top \eta\right)} \right) \right] \tag{3.14}$$

3.2. Estimación e Inferencia

El estimador de máxima verosimilitud $\hat{\theta}$ para θ viene dado por:

$$\hat{\theta} = \arg\max_{\gamma,\beta,\eta} \ell(\theta|\Gamma) \tag{3.15}$$

Como se observa, la log-verosimilitud del modelo en (3.14) es bastante compleja y por lo tanto no hay una solución explicita. En este caso se debe utilizar métodos numéricos para hallar los estimadores de máxima verosimilitud de los parámetros como el método BFGS a partir de la función de log-verosimilitud (3.14) con $\kappa_l = \log(\gamma_j),\ l = 1 \dots k$. Sin embargo, en el presente trabajo se utilizará el algoritmo de Esperanza y Maximización (EM) para hallar los estimadores (Dempster et al., 1977).

3.2.1. Algoritmo EM

En el modelo propuesto, la variable que indica si la observación es suceptible al evento de ínteres, Y, es parcialmente observada por lo que la consideramos latente. Con esta premisa, es posible utilizar el algoritmo EM para estimar los parámetros del modelo.

Paso E: Esperanza

En este paso definimos $Q(\theta|\theta^{(t)})$ como la esperanza de la log-verosimilitud de $\theta = \{\gamma, \beta, \eta\}$ con respecto a la distribución de y_i dado Γ y los estimados actuales de $\theta^{(t)} = \left\{\gamma^{(t)}, \beta^{(t)}, \eta^{(t)}\right\}$:

$$Q(\theta|\theta^{(t)}) = E_{y_i|\Gamma,\theta^{(t)}}\left[\log \mathcal{L}\left(\theta; \Gamma, Y\right)\right] \tag{3.16}$$

Para realizar este cálculo es necesario conocer la función de verosimiltud de los datos completos $\mathcal{L}\left(\theta; \Gamma, Y\right)$. Dado que conocemos los valores de π_i, este viene dado por

$$\mathcal{L}\left(\theta; \Gamma, Y\right) = \prod_{i=1}^{n} \left(P(\theta|y_i = 1, \Gamma_i, \delta_i = 0)\right)^{I(y_i=1)I(\delta_i=0)}$$
$$\left(P(\theta|y_i = 0, \Gamma_i, \delta_i = 0)\right)^{I(y_i=0)I(\delta_i=0)}$$
$$\left(P(\theta|y_i = 1, \Gamma, \delta_i = 1)\right)^{I(y_i=1)I(\delta_i=1)} \tag{3.17}$$

Cabe resaltar que no se considera el término $\left(P(\theta|y_i = 0, \Gamma, \delta_i = 1)\right)^{I(y_i=0)I(\delta_i=1)}$ ya que $I(y_i = 0)I(\delta_i = 1) = 0$ para todo i. Es decir, no existe la posibilidad de que aparezca el evento y que el sujeto

de estudio pertenezca a la fracción curada.

Por su parte se tiene que

$$P(\theta|y_i = 1, \Gamma_i, \delta_i = 0) = \frac{\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}{1 + \exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})} \tag{3.18}$$

$$P(\theta|y_i = 0, \Gamma_i, \delta_i = 0) = \frac{\exp\left(-\sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\right)}{1 + \exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})} \tag{3.19}$$

$$P(\theta|y_i = 1, \Gamma_i, \delta_i = 1) = \frac{\exp\left(-\sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\right) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}) \sum_{l=1}^{k} \gamma_j \boldsymbol{M}_j(t)}{1 + \exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}. \tag{3.20}$$

Por lo tanto,

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E_{Y|\Gamma,\theta^{(t)}}\left[\log \mathcal{L}\left(\theta; \Gamma, Y\right)\right] \\ &= E_{Y|\Gamma,\theta^{(t)}}\left[\log \prod_{i=1}^{n} \mathcal{L}\left(\theta; \Gamma_i, y_i\right)\right] \\ &= E_{Y|\Gamma,\theta^{(t)}}\left[\sum_{i=1}^{n} \log \mathcal{L}\left(\theta; \Gamma_i, y_i\right)\right] \\ &= \sum_{i=1}^{n} E_{y_i|\Gamma,\theta^{(t)}}\left[\log \mathcal{L}\left(\theta; \Gamma_i, y_i\right)\right]. \end{aligned} \tag{3.21}$$

Lo que nos lleva a que

$$\begin{aligned} Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} &\log \mathcal{L}\left(\theta; \Gamma_i, y_i = 0\right) P(y_i = 0|\Gamma_i, \theta^{(t)}) \\ &+ \log \mathcal{L}\left(\theta; \Gamma_i, y_i = 1\right) P(y_i = 1|\Gamma_i, \theta^{(t)}) \end{aligned} \tag{3.22}$$

A partir de (3.18) y (3.19), se tiene que

$$P(y_i = 0|\Gamma_i, \theta^{(t)}, \delta_i = 0) = \frac{\frac{\exp\left(-\sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\right)}{1+\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}}{\frac{\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}{1+\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})} + \frac{\exp\left(-\sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\right)}{1+\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}}. \tag{3.23}$$

Para simplificar la notación se define $\Pi_{1i} = P(y_i = 1|\Gamma_i, \theta^{(t)}, \delta_i = 0) = 1 - P(y_i = 0|\Gamma_i, \theta^{(t)}, \delta_i = 0)$. En este sentido, se tiene que

$$
\begin{aligned}
\Pi_{1i} &= P(y_i = 1|\Gamma_i, \theta^{(t)}, \delta_i = 0) \\
&= 1 - P(y_i = 0|\Gamma_i, \theta^{(t)}, \delta_i = 0) \\
&= \frac{\frac{\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}{1+\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}}{\frac{\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}{1+\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})} + \frac{\exp\left(-\sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\right)}{1+\exp(-\boldsymbol{z_i}^\top \boldsymbol{\eta})}}
\end{aligned}
\tag{3.24}
$$

Por último, luego de reemplazar (3.17), (3.18), (3.19) y (3.20) en (3.22) se sigue que

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) = &\sum_{i=1}^{n} \Pi_{1i} I(\delta_i = 0) \left(-\boldsymbol{Z}^\top \eta - \log\left(1 + \exp\left(-\boldsymbol{Z}^\top \eta\right)\right)\right) + \\
&\left(1 - \Pi_{1i}\right) I(\delta_i = 0) \left(- \sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}) - \log\left(1 + \exp\left(-\boldsymbol{Z}^\top \eta\right)\right) \right) + \\
&I(\delta_i = 1) \left(- \sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}) + \boldsymbol{x_i}^\top \boldsymbol{\beta} \right) + \\
&I(\delta_i = 1) \left(\log\left(\sum_{l=1}^{k} \gamma_j \boldsymbol{M}_j(t) \right) - \log\left(1 + \exp\left(-\boldsymbol{Z}^\top \eta\right)\right) \right) \\
= &\sum_{i=1}^{n} - I(\delta_i = 0) \boldsymbol{Z}^\top \eta \Pi_{1i} - \log\left(1 + \exp\left(-\boldsymbol{Z}^\top \eta\right)\right) + \\
&\sum_{i=1}^{n} \sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}) \left(\Pi_{1i} I(\delta_i = 0) - 1\right) + \\
&I(\delta_i = 1) \left(\boldsymbol{x_i}^\top \boldsymbol{\beta} + \log\left(\sum_{l=1}^{k} \gamma_j \boldsymbol{M}_j(t) \right) \right)
\end{aligned}
\tag{3.25}
$$

Paso M: Maximización

En este paso, se procede a actualizar θ de la siguiente forma:

$$
\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)}) \tag{3.26}
$$

Como se puede observar, la maximización de $Q(\theta \mid \theta^{(t)})$ es separable en la maximización de $Q(\eta \mid \theta^{(t)})$ y de $Q(\gamma, \beta \mid \theta^{(t)})$ definidos como

$$
Q(\eta \mid \theta^{(t)}) = \sum_{i=1}^{n} -I(\delta_i = 0) \boldsymbol{Z}^\top \eta \Pi_1 - \log\left(1 + \exp\left(-\boldsymbol{Z}^\top \eta\right)\right) \tag{3.27}
$$

$$
\begin{aligned}
Q(\gamma, \beta \mid \theta^{(t)}) = &\sum_{i=1}^{n} \sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t) \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}) \left(\Pi_1 I(\delta_i = 0) - 1\right) \\
&+ I(\delta_i = 1) \left(\boldsymbol{x_i}^\top \boldsymbol{\beta} + \log\left(\sum_{l=1}^{k} \gamma_j \boldsymbol{M}_j(t) \right) \right).
\end{aligned}
\tag{3.28}
$$

Más aún, se puede ver que (3.27) tiene la forma de la verosimilitud de un modelo binomial con covariables $\boldsymbol{Z}$ y variable dependiente $\Pi_1 \cdot I(\delta_i = 0)$. Por otro lado, para maximizar $Q(\gamma, \beta \mid \theta^{(t)})$ se opta por reemplazar γ por $\kappa = \log(\gamma)$, ya que ayuda a su convergencia y continuidad de la función, y de esta manera no se tiene que restringir al parámetro a ningún dominio. Asimismo, se utilizará el método BFGS en conjunto con su gradiente compuesta por:

$$\begin{aligned}\frac{\partial Q}{\partial \kappa_l} =& \frac{\partial Q}{\partial \gamma_j}\frac{\partial \gamma_j}{\partial \kappa_l} \\ =& \sum_{i=1}^{n} \exp(\kappa_l)\boldsymbol{I}_j(t)\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\left(\Pi_1 I(\delta_i = 0) - 1\right) + \\ & \sum_{i=1}^{n} \exp(\kappa_l)\frac{I(\delta_i = 1)\boldsymbol{M}_j(t)}{\sum_{j=1}^{k}\exp(\kappa_j)\boldsymbol{M}_j(t)}\end{aligned} \tag{3.29}$$

$$\frac{\partial Q}{\partial \beta_l} = \sum_{i=1}^{n}\sum_{j=1}^{k} \exp(\kappa_j)\boldsymbol{I}_j(t)x_l\exp(\boldsymbol{x_i}^\top\boldsymbol{\beta})\left(\Pi_1 I(\delta_i = 0) - 1\right) + I(\delta_i = 1)x_l \tag{3.30}$$

En resumen, para aplicar el algoritmo realizamos los siguientes pasos:

I) Incializar los parámetros θ con valores aleatorios

II) Calcular Π_1 dado los valores de θ actuales

III) Actualizar los valores de θ según el paso M

IV) Iterar los pasos II y III hasta convergencia

3.3. Estimación de la varianza

Para evaluar el ajuste de los estimadores de los parámetros del modelo estimaremos la varianza de los mismos. Para esto haremos uso del estimador de la varianza del método SEM Cai and Lee (2009). Cabe resaltar que es posible hallar una aproximación a partir de la hessiana de la maximización de la log-verosimilitud si se hubiese utilizado el método BFGS.

El algoritmo SEM o Supplemented EM permite el uso de los pasos del algoritmo EM para estimar la matriz de varianza-covarianza de los parámetros hallados. Este hace uso del principio de información faltante Orchard and Woodbury (1972) como lo expresa Cai (2008), que indica la información total $\mathcal{I}(\hat{\theta}|Y_{obs})$ en función de la información de los datos completos $\mathcal{I}_c(\hat{\theta})$ y la información de los datos faltantes $\mathcal{I}_m(\hat{\theta})$

$$\mathcal{I}(\hat{\theta}|Y_{obs}) = \mathcal{I}_c(\hat{\theta}) - \mathcal{I}_m(\hat{\theta}) \tag{3.31}$$

Este algoritmo utiliza el algoritmo EM como un mapeo Φ de θ, un subconjunto de d dimensiones de $\mathbb{R}^d$, de manera que en la iteración k se tiene que:

$$\theta^{(k+1)} = \Phi(\theta^{(k)}) \tag{3.32}$$

Según la expansión de la serie de Taylor para $\Phi(\theta)$, cerca de la vecindad de $\hat{\theta}$ donde $\Phi(\hat{\theta}) = \hat{\theta}$, se tiene la aproximación:

$$\theta^{(k+1)} \approx \hat{\theta} + \Delta(\hat{\theta})(\theta^{(k)} - \hat{\theta}) \tag{3.33}$$

donde $\Delta(\hat{\theta})$ es la matriz Jacobiana $d \times d$ de $\Phi(\theta)$.

Dempster et al. (1977) muestra que en la vecindad de $\hat{\theta}$ se tiene que

$$\Delta(\hat{\theta}) = \mathcal{I}_m(\hat{\theta})\mathcal{I}_c^{-1}(\hat{\theta}), \tag{3.34}$$

lo cual, según ciertas condiciones explicadas en Cai (2008), se puede invertir en

$$V(\hat{\theta}|Y_{obs}) = \mathcal{I}^{-1}(\hat{\theta}|Y_{obs}) = \mathcal{I}_c^{-1}(\hat{\theta})\left\{I_d - \Delta(\hat{\theta})\right\}^{-1}, \tag{3.35}$$

donde I_d es la matriz identidad $d \times d$.

Para el cálculo de $\mathcal{I}_c(\hat{\theta})$ aplicado al modelo propuesto se tiene

$$\mathcal{I}_{c,\gamma_j\gamma_m}(\hat{\theta}) = -\sum_{i=1}^{n} \frac{I(\delta_i = 1)\boldsymbol{M}_j \boldsymbol{M}_m}{\left(\sum_{l=1}^{k} \gamma_j \boldsymbol{M}_j(t)\right)^2} \tag{3.36}$$

$$\mathcal{I}_{c,\eta_j\eta_m}(\hat{\theta}) = -\sum_{i=1}^{n} \frac{Z_j Z_m \exp(-Z^\top \eta)}{\left(\exp(-Z^\top \eta) + 1\right)^2} \tag{3.37}$$

$$\mathcal{I}_{c,\beta_j\beta_m}(\hat{\theta}) = \sum_{i=1}^{n} X_j X_m \exp(X^\top \beta)\left(\sum_{l=1}^{k} \gamma_j \boldsymbol{I}_j(t)\right)\left(\Pi_1 I(\delta_i = 0) - 1\right) \tag{3.38}$$

$$\mathcal{I}_{c,\beta_j\gamma_m}(\hat{\theta}) = \sum_{i=1}^{n} X_j \exp(X^\top \beta) I_k \left(\Pi_1 I(\delta_i = 0) - 1\right). \tag{3.39}$$

Por su parte, cada componente de $\Delta(\hat{\theta})$ definido como r_{ij}, puede ser aproximado con el siguiente algoritmo.

Algoritmo de estimación de $\Delta(\hat{\theta})$

I) Definir el vector $\theta_{(i)}^{(k)} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_{i-1}, \theta_i^{(k)}, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_d\right)$

II) Correr una iteración del mapeo $\theta_{(i)}^{(t+1)} = \Phi(\theta_{(i)}^{(t)})$

III) Obtener el ratio $r_{ij} = \frac{\theta_{j(i)}^{(t+1)} - \hat{\theta}_j}{\theta_i^{(t)} - \hat{\theta}_i}$, para $j = 1, \ldots, d$.

IV) Iterar los pasos I, II y III hasta convergencia de cada r_{ij}

Capítulo 4

Estudio de Simulación

Para estudiar el proceso de estimación presentado, en determinados escenarios, se procede a simular datos según una estructura requerida y probar la recuperación de los parámetros preestablecidos.

4.1. Simulador de Datos

Para la simulación de los datos requeridos se necesita establecer varios parámetros e hiperparámetros. Entre estos está el tiempo límite del experimento sobre el cual todos los tiempos corresponden a censuras. Asimismo, se tienen los parámetros de los splines como los nodos correspondientes, su grado respectivo de continuidad y los coeficientes asociados. De las características de las observaciones hay que definir el número de covariables y sus distribuciones que influyen en la función de riesgo de manera proporcional o sobre la probabilidad de no pertenecer al grupo de riesgo. Por último, se define la distribución del tiempo de censura el cual limita cuantas veces el evento deseado se observa.

El Cuadro 4.1 muestra los parámetros supuestos para los casos desarrollados

	Condición	Valor
1)	Tiempo límite del experimento	10
2)	Nodos de splines	$\{2, 4, 6\}$
3)	Grado de splines	3 (cúbico)
4)	Número de covariables en $\boldsymbol{X}$	4
5)	Número de covariables en $\boldsymbol{Z}$	intercepto + 2 covariables
6)	Coeficientes de $\boldsymbol{X}$	$\{-1, -0{,}25, 0{,}25, 1\}$
7)	Coeficientes de $\boldsymbol{Z}$	$\{1, -0{,}5, 0{,}5\}$
8)	Coeficientes de splines	$\{0{,}3, 0{,}6, 0{,}9, 1{,}2, 0{,}6, 0{,}15\}$
9)	Distribución del tiempo de censura Y	Uniforme(0, 10)
10)	Distribución de covariables $\boldsymbol{X}$ y $\boldsymbol{Z}$	Normal Truncada($\mu = 0, \sigma = 1$, mín $= -1{,}5$, máx $= 1{,}5$)

Cuadro 4.1: Condiciones para la simulación de datos

Una de las mayores dificultades es la simulación de los tiempos del evento, ya que estos dependen de los splines los cuales no tienen una ecuación explícita. Es por esto que no es posible utilizar el método de la transformada inversa. Para esta aplicación se utiliza una versión modificada del algoritmo de Newton-Raphson (Anexo 7.1 para mayores detalles) para rangos limitados de búsqueda donde se genera una variable aleatoria uniforme U entre 0 y 1 y se halla el valor de t donde $S(t) - U = 0$ y, en caso de divergencia, el método propuesto en el Anexo 7.2 y se simulan los tiempos para el evento de abandono.

4.2. Ejemplo: Base de datos simulada

Para ilustrar el proceso de simulación, se generará una sola base de datos de 10 000 observaciones. En este caso se tiene un gran grupo de observaciones con $t = 10$. Esto se debe a que en muchos casos la probabilidad acumulada hasta $t = 10$ es menor que 1. No obstante, esto no representa un problema ya que se estableció previamente que el experimento dura máximo 10 unidades de tiempo y cualquier $t > 10$ será por definición una censura. La Figura 4.1 ilustra las frecuencias de los datos simulados (barras) y la densidad teórica del tiempo que se quiere simular (curva roja).

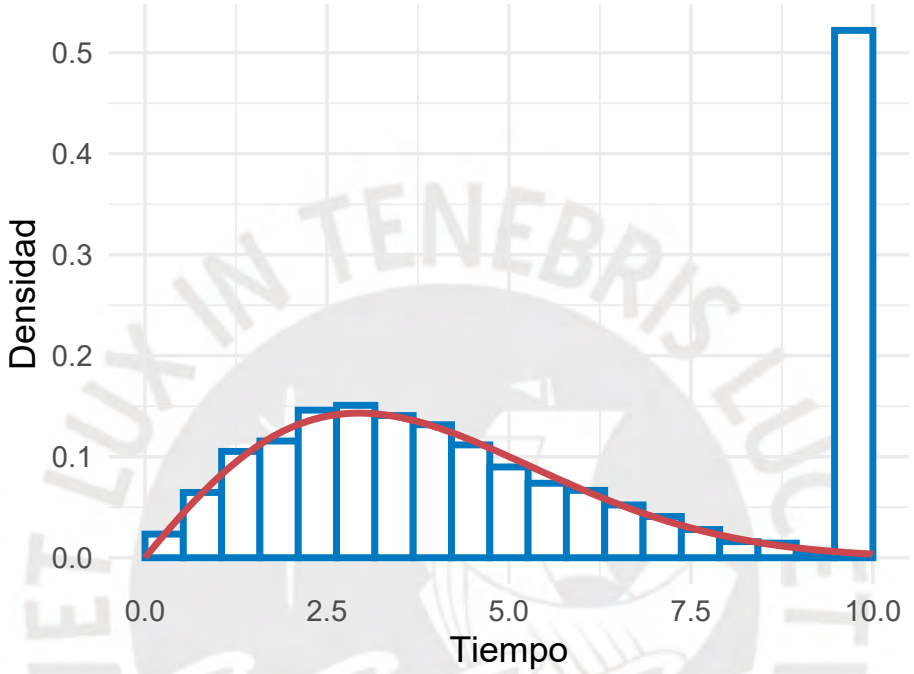


Figura 4.1: Distribución de los tiempos simulados por el proceso propuesto

En la implementación se desea recuperar $\theta = \{\gamma, \beta, \eta\}$ dado un conjunto de datos simulado. Todos los valores de θ son inicializados en 0.5, sin embargo esta elección sera discutida en la sección 4.3. De esta manera, se utiliza el algoritmo EM y como condición de convergencia se establece que la log-verosimilitud mejore en menos de 0.0001 %.

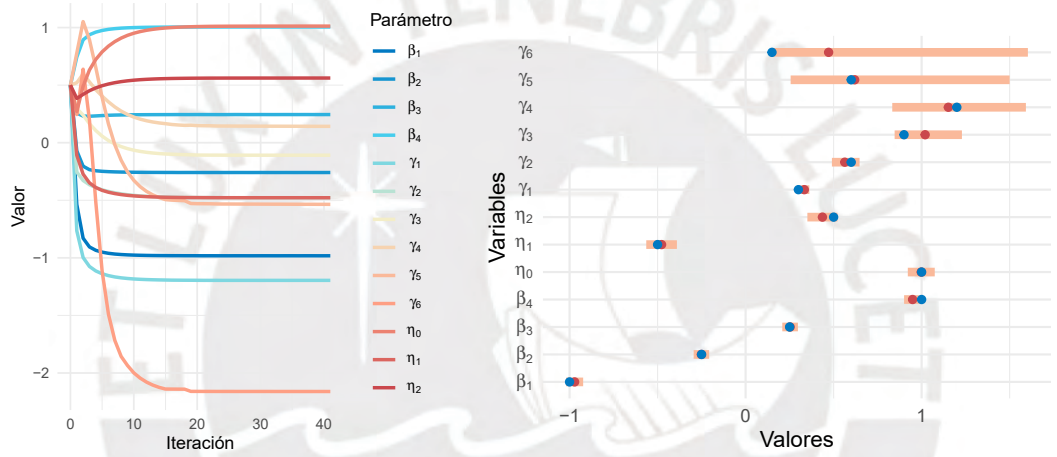
Para esta base de datos en particular, el algoritmo converge luego de 20 iteraciones y las estimaciones se pueden observar en el Cuadro 4.2. La Figura 4.2a muestra las iteraciones hasta la convergencia de $\hat{\theta}$ dado los datos simulados. Asimismo, la Figura 4.2b presenta los valores estimados y sus intervalos de confianza asintóticos a un nivel de confianza del 95 % donde los puntos rojos son los valores estimados y los azules son los simulados. Como se observa, los parámetros son recuperados de manera satisfactoria y con gran precisión. Sin embargo, γ_5 y γ_6 no siempre tienen un valor significativamente mayor a cero. Esto se debe en gran parte a que su aporte original a la función de riesgo es pequeño y su efecto se da en la cola derecha, donde la cantidad de datos con la que se cuenta es escasa y da poca confiabilidad.

4.3. Pruebas de Estabilidad

Para validar la estabilidad del modelo se procede a simular 1000 casos con tres escenarios distintos para los parámetros y tres escenarios para el número de observaciones. Los resultados de estas

Parámetro	Estimado	Real
γ_1	0.30	0.30
γ_2	0.62	0.60
γ_3	0.91	0.90
γ_4	1.22	1.20
γ_5	0.67	0.60
γ_6	0.14	0.15
η_0	1.00	1.00
η_1	-0.47	-0.50
η_2	0.56	0.50
β_1	-0.98	-1.00
β_2	-0.26	-0.25
β_3	0.24	0.25
β_4	1.01	1.00

Cuadro 4.2: Resultado de los estimadores de los datos simulados: Un solo conjunto de datos



(a) Convergencia de estimadores por el algoritmo EM

(b) Intervalos de confianza de los estimadores

Figura 4.2: Recuperación de parámetros de datos simulados: Un solo conjunto de datos

simulaciones se pueden observar en el Cuadro 4.3.

En estos resultados se puede observar todos los parámetros se pueden recuperar adecuadamente con la expeción de γ_3. Fuera de esa excepción, las simulaciones muestran que el estimador es insesgado y que la cobertura es la apropiada. En estos casos, la conclusion es la misma a lo largo de los escenarios de parámetros y diferentes tamanõs de muestra considerados.

En el caso particular de γ_3, esto se puede deber, como se explicó anteriormente, a que para la cola de la función de riesgo no se tienen muchos datos y por ende la precisión disminuye y por ende la estimación se hace menos estable.

Parámetro	n = 250		n = 500		n = 1000	
	Sesgo	Cobertura	Sesgo	Cobertura	Sesgo	Cobertura
Escenario 1						
$\beta_1 = 2$	0.018	0.935	0.01	0.924	0.004	0.941
$\eta_0 = 2$	0.025	0.966	0.01	0.964	-0.003	0.958
$\eta_1 = 0.5$	0.053	0.952	-0.018	0.956	-0.033	0.962
$\gamma_1 = 0.5$	0.072	0.95	0.017	0.954	0.015	0.951
$\gamma_2 = 2$	-0.098	0.967	-0.076	0.963	-0.03	0.967
$\gamma_3 = 0.5$	0.149	0.857	-0.002	0.876	0.027	0.901
Escenario 2						
$\beta_1 = 1.5$	0.005	0.939	0.004	0.94	-0.011	0.915
$\eta_0 = 2.5$	0.016	0.969	0.009	0.964	-0.001	0.936
$\eta_1 =$ -0.5	0.041	0.948	-0.022	0.946	-0.01	0.954
$\gamma_1 = 6$	-0.035	0.883	-0.026	0.936	0.02	0.77
$\gamma_2 = 4$	-2.266	0.936	-0.072	0.925	-3.8	0.954
$\gamma_3 = 10$	0.075	0.836	-0.003	0.88	0.219	0.658
Escenario 3						
$\beta_1 = -2$	-0.01	0.936	0.001	0.941	-0.018	0.854
$\eta_0 = 3$	0.025	0.96	0.003	0.95	0.006	0.956
$\eta_1 = -1$	0.024	0.944	0.008	0.953	0.008	0.947
$\gamma_1 = 8$	-0.01	0.874	-0.014	0.923	0.016	0.692
$\gamma_2 = 5$	-3.175	0.951	-0.167	0.949	-3.822	0.947
$\gamma_3 = 0.1$	-1.039	0.593	0.428	0.707	-1.037	0.567

Cuadro 4.3: Resultado de las pruebas de estabilidad en base a 1000 bases de datos generadas en cada escenario

Capítulo 5

Aplicación

En este capítulo se discutirá la aplicación del modelo de larga duración estudiado en los capítulos anteriores usando datos de una de las más grandes aseguradoras del Perú. Por temas de seguridad, los datos han sido cifrados y las variables continuas normalizadas.

Se utilizó 10 años de información observada de un solo producto de la empresa. Para este producto se consideraron todos lo clientes desde su fecha de emisión del producto hasta su anulación o fecha de corte de los datos que corresponde al 29 de Noviembre de 2018. Se consideran inclusive las pólizas cercanas a la fechas de corte ya que hay gran cantidad de pólizas que son anuladas al principio de su vigencia.

El evento de interés será la cancelación o caducidad de la póliza. Es así que una persona que adquirió el producto el 4 de Octubre de 2016, y este sigue vigente hasta la última fecha de observación, tiene un tiempo observado de 786 días y es categorizado como dato censurado ya que el evento no ocurrió. En este sentido, el foco sobre el cual se concentra la investigación es el tiempo hasta la cancelación del producto y la fracción de clientes que son inmunes a este evento. Tal como se expresó, el tiempo observado puede ser censurado y esto se puede dar por tres motivos: i) todavía no ha pasado suficiente tiempo para que el cliente abandone el producto, ii) los clientes que incurren en un siniestro que consume el producto y se considera el momento de este siniestro como el tiempo observado y censurado y iii) la observación puede estar censurada porque el individuo pertenece a la fracción de la población que es inmune al riesgo (aun cuando esto último no es distinguible en nuestros datos).

El objetivo de esta capítulo será evaluar qué factores influyen sobre la probabilidad de ser inmunes al abandono y sobre el tiempo hasta el abandono para nos susceptibles. Entre las posibles variables a considerar están la edad de la persona, el costo anual del producto, la frecuencia de pago, el sexo de la persona, su estado civil, el número de hijos registrados y si ha tenido problemas de salud anteriormente.

5.1. Estructura de Datos

Se tiene información de las pólizas emitidas entre las cuales se encuentra la fecha de emisión de la póliza. Esta información se obtuvo a una fecha de corte determinada. El evento que queremos observar es la cancelación de la póliza, pero requiere que sea construido.

La Tabla 5.1 muestra la estructura de datos que se tiene a disposición. Nótese que `ID` es el identificador del cliente, `x1`, `x2` hasta `xn` son las covariables con las que se dispone, `Fecha Emisión` es la fecha en la que inicia la póliza, `Estado` contiene la información de si la póliza ha sido cancelada o no y `Fecha Estado` es la fecha de la observación del registro. Finalmente, a partir de esta información se crea el evento donde el `Estado` de la póliza es `Cancelada` y la `Fecha Estado` sea menor que

la fecha de corte. Todos los demás se consideran censuras por derecha. Además, se crea el tiempo observado que para los casos en los que ocurrió el evento viene dado como el número de días desde `Fecha Emisión` hasta `Fecha Estado` y para los que no ocurrió, se define como el número de días desde `Fecha Emisión` hasta la fecha de corte o la `Fecha Estado` para la pólizas siniestradas.

ID	x1	x2	...	xn	Fecha Emisión	Estado	Fecha Estado
1	x11	x12	...	x1n	1990-01-05	Vigente	-
2	x21	x22	...	x2n	1998-06-25	Cancelada	2000-01-29
...	...	...	...	...	...	...	...
m	xm1	xm2	...	xmn	1995	Cancelada	1998-09-18

Cuadro 5.1: Ejemplo de la estructura de los datos

El Cuadro 5.2 presenta las principales medidas de resumen de las variables consideradas y que podrían estar asociadas con la cancelación o no de pólizas. Un poco menos de la mitad de la muestra (46.04 %) reportó un abandono de su póliza. De igual manera, las personas que abandonan su póliza tienden a ser menores (medias: 38 vs. 40 años), ser en un mayor porcentaje mujeres (39 % vs. 37 %) y tener menores problemas de salud (2.17 % vs. 9.27 %) en comparación con personas que no abandonaron su póliza.

Variables	Censurado 47,579 (53.96 %)	No Censurado 10,594 (46.04 %)	Total
Edad anõs (media)	40.00	37.54	38.87
Prima anualizada USD (media)	178.32	179.18	178.72
Indicador de frecuencia no mensual de pago	2.72 %	1.30 %	2.17 %
Número de hijos registrados del cliente (media)	1.03	0.93	0.98
Indicador de problemas de salud registrados	9.27 %	2.17 %	6 %
Sexo del cliente			
Femenino	37 %	39 %	38 %
Masculino	63 %	61 %	62 %
Estado civil del cliente			
Casado	49.46 %	36.45 %	43.47 %
Soltero	42.49 %	55.99 %	48.70 %
Otros	8.05 %	7.56 %	7.83 %

Cuadro 5.2: Características de la muestra de clientes de la empresa aseguradora ($n = 58{,}173$)

5.2. Modelo de Larga Duración sin Covariables

En la Figura 5.1 se muestran las estimaciones de los parámetros del modelo sin covariables descrito previamente (3.11). Esta estimación es constrastada con el estimador no paramétrico de Kaplan-Meier (en color rojo). Se puede observar que el estimador Kaplan-Meier tiene una forma algo irregular que difícilmente puede ser capturado por un modelo paramétrico sencillo. Sin embargo, dada la naturaleza semiparamétrica de los splines es posible definir los suficientes nodos para que el estimador sea lo suficientemente flexible para modelar el comportamiento observado de los datos. En la Figura 5.1a se muestra el estimador del modelo propuesto considerando solo un nodo. Asimismo, es posible definir las splines con cuatro nodos como (Figura 5.1b). Note que esta función tiene un mejor ajuste al estimador Kaplan-Meier y, como se verá posteriormente, un mejor AIC sustentado por la cantidad de observaciones con las que se cuenta (Cuadro 5.3).

Adicionalmente, esto se evidencia en el gráfico 5.1c y 5.1d que presenta el ajuste de un modelo de larga duración Gamma y otro Weibull. Se puede observar que hay importantes diferencias con respecto al estimador Kaplan-Meier. Más aún, como se observa en el Cuadro 5.3, el AIC y BIC del modelo con splines es menor, lo que indica un mejor ajuste. En este caso de uso con los datos presentados, no se analiza el AICc ya que la muestra es suficientemente grande para el número de parámetros que se están utilizando. Cabe mencionar que en el modelo sin convariable se obtiene un estimado de la fracción de cura de 17 % con un intervalo de confianza al 95 % de [8 % - 34 %]. Este intervalo, a pesar de ser significativamente distinto a cero, es basta amplio. Esto se puede deber a que la función de supervivencia en el tiempo máximo no parece haberse estabilizado.

Figura 5.1: Función de supervivencia estimada por distintos métodos: a) con un nodo, b) con cuatro nodos, c) con una distribución gamma y d) con una distribución Weibull

Modelo	AIC	BIC
Gamma	721214.3	721242.5
Weibull	721347.7	721375.9
Splines con 1 nodo	720062.0	720055.5
Splines con 4 nodos	717424.3	717413.9

Cuadro 5.3: Ajuste relativo de modelos a los datos reales

5.3. Modelo de Larga Duración con Covariables

En esta sección se construyó un modelo de regresión para medir el impacto de las covariables sobre la probabilidad de ser susceptible o no y sobre el tiempo a la ocurrencia del evento de ínteres entre los susceptibles. En este caso se consideraron varias covariables, sin embargo solo se mantuvieron las que tenían significancia estadística (un valor de p menor a 0.05). En este sentido, se usó la

variable `Indicador de problemas de salud registrados` para modelar la probabilidad de ser susceptible al evento de interés y las potenciales variables asociadas con el riesgo de la ocurrencia del evento fueron la prima anualizada estandarizada, la edad estandarizada y el indicador de la frecuencia de pago mensual.

	Estimación	Límite Inferior	Límite Superior
Coeficientes de Spline			
gamma_1	0.01	0.01	0.01
gamma_2	0.10	0.08	0.12
gamma_3	0.02	0.01	0.02
gamma_4	0.25	0.18	0.34
gamma_5	0.07	0.05	0.10
gamma_6	0.13	0.09	0.19
gamma_7	0.19	0.13	0.27
Modelo Proporcional			
xh_prima	0.14	0.14	0.15
xh_mensual	0.92	0.25	1.59
xh_edad	-0.28	-0.29	-0.27
Modelo Logit			
(Intercept)	-2.23	-2.31	-2.15
xp_ind_salud	1.51	1.39	1.63

Cuadro 5.4: Resultado de los estimadores de los datos reales

El cuadro 5.4 muestra los resultados de dicho modelo. Se puede corroborar la hipótesis de que las personas que presentaron algún problema de salud anteriormente son menos propensos a abandonar su seguro de vida. Asimismo, el riesgo de abandono aumenta con el aumento de la prima, si es que lo pagos se realizan de manera mensual y disminuyen con el aumento de la edad de adquisición de la póliza. En específico, si la persona ha tenido algún problema de salud, la probabilidad de pertenecer a la fracción de la población que no participa en el proceso de abandono aumenta en 23 %. De igual manera, el riesgo de abandono de la póliza aumenta en aproximadamente 2.5 veces si esta se paga mensualmente. No obstante, dado su intervalo de confianza, el mínimo que aumenta el riesgo es de 1.3 veces y el máximo es de 4.9 veces.

El cuadro 5.5 muestra las estimaciones usando el modelo propuesto y un modelo Gamma de larga duración considerando las mismas covariables. Se puede observar que el efecto de las variables utilizadas tienen un comportamiento parecido, sin embargo existen diferencias sustantivas que pueden aparecer dependiendo del tipo de modelo que se utilice. En específico, se puede observar que el modelo Gamma estaría sobreestimando el efecto del indicador de problemas de salud. De la misma manera, el efecto del pago mensual sobre el riesgo parece menor en el modelo Gamma; sin embargo, no es significativamente distinto dado los intervalos de confianza.

Se puede ver que el modelo propuesto con splines monótonos tiene un mejor ajuste y que el impacto de las variables a analizar presenta variaciones dependiendo del modelo utilizado (Cuadro 5.6). Más aún, la forma de la función de riesgo es distinta entre un modelo y otro, lo que puede llevar a una interpretación incorrecta de los resultados.

		Modelo Splines		Modelo Gamma	
		Estimado	Error estándar	Estimado	Error Estándar
M. Logit	xp_ind_salud	1.51	0.06	1.87	0.04
M.R.P.	xh_prima	0.14	0.00	0.13	0.00
	xh_mensual	0.92	0.34	0.86	0.05
	xh_edad	-0.28	0.00	-0.29	0.01

Cuadro 5.5: Comparación de estimadores de los datos reales según modelo propuesto vs. modelo Gamma

	Modelo Splines	Modelo Gamma
AIC	712708.7	716886.2
BIC	712821.4	716951.9

Cuadro 5.6: Desempeño del modelo propuesto vs. modelo Gamma

Capítulo 6

Conclusiones y Discusión Final

6.1. Conclusiones

En esta investigación se ha trabajado el desarrollo y aplicación de un modelo de supervivencia de larga duración donde el tiempo a desarrollar el evento de interés, entre los susceptibles, satisface el supuesto de riesgos proporcionales y su función de riesgo basal se construye usando splines monótonos.

En ese sentido, entre las características más importantes del modelo propuesto está la posibilidad de evaluar factores asociados con el ser suceptible o no al evento de interes y de evaluar factores asociados riesgo de acelerar o retrasar la ocurrencia del evento entre los susceptibles.

En el desarrollo de las simulaciones se observó que el modelo recupera los parámetros de manera correcta y consistente. Sin embargo, puede que los parámetros correspondientes a los splines tengan mucha varianza para sus componentes de tiempo alto. Esto se puede deber a que para tiempos altos donde está la importancia de estas componentes, se tienen muy pocos datos, ya sea por censura o por eventos ocurridos anteriormente. Adicionalmente, pudimos corroborar que el modelo propuesto logra capturar funciones de supervivencia con formas no regulares que otros modelos paramétricos, como el de larga duración gamma, no logran modelar.

Finalmente, se aplico la metodología estudiada a un conjunto de datos reales de una empresa de seguros donde el evento de interés se definió como el abandono del seguro. En este análisis fue posible identificar que el mayor valor de prima y la facturación mensual de la póliza aumentan el riesgo de cancelación de la póliza, mientras que el aumento de la edad del contratante disminuye el riesgo. De igual forma, se identificó que el haber tenido problema de salud disminuye en gran manera la probabilidad de pertenecer a la fracción de la población que es susceptible al abandono.

Por otro lado, la estimación propuesta con el algoritmo EM para hallar los EMV de los parámetros requeridos es bastante eficiente. En resumen, el modelo propuesto es eficiente computacionalmente y bastante flexible sobre todo en el caso de estimación de la función basal lo cual permite que este modelo se pueda aplicar a diversos escenarios.

6.2. Investigaciones futuras

1. Extender el modelo propuesto para distintos tipos de censura como por ejemplo censura por intervalo.

2. Desarrollar un método eficiente para estimar el número de nodos óptimos para los splines monótonos.

3. Desarrollar la estimación bayesiana para el modelo propuesto.

Capítulo 7

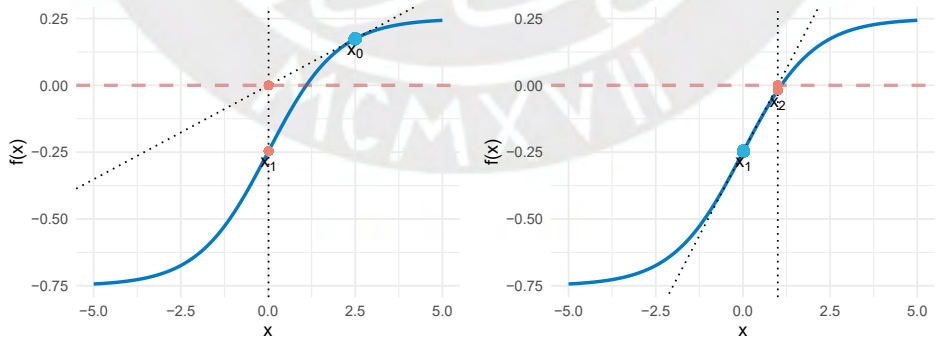
Apendices

7.1. Método de Newton-Raphson

El método de Newton-Raphson es un algoritmo que sirve para hallar la raíz de una función f; esto es, halla x tal que $f(x) = 0$. Es un proceso iterativo y depende de ciertas condiciones para su convergencia. En la presente investigación se utilizará este método para la simulación de los datos de tiempo a partir de una función semiparamétrica. El algoritmo va de la siguiente forma:

1. Seleccionar un punto de partida x_0
2. Hallar $x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$
3. Repetir el paso 2 hasta que $x_{t+1} = x_t$ o su variación sea despreciable.

Una representación gráfica de las dos primeras iteraciones se puede observar en la Figura 7.1. Como se menciona, el punto inicial debe estar suficientemente cercano a la raíz para que el algoritmo pueda converger ya que los puntos de inflexión pueden causar que los valores hallados en cada iteración se alejen más de la raíz lo que puede resultar en falta de covergencia. En el ejemplo dado en la Figura 7.1, si el punto inicial x_0 se daba en $-2{,}5$ por dar un ejemplo, se puede ver que el algoritmo diverge iterando finalmente entre $+\infty$ y $-\infty$.



(a) Primera Iteración (b) Segunda Iteración

Figura 7.1: Método de Newton-Raphson

7.2. Simulación de Números Aleatorios

Para el desarrollo de esta investigación será necesario simular ciertos datos a partir de su distribución. Sin embargo, muchas veces no contamos con su expresión analítica por lo que restringe el

método que se utilizará. En este caso, no se cuenta con la expresión de su función de probabilidad acumulada inversa ni tampoco se puede probar que sea log-concava por lo que se optó por el siguiente método para simular t dada la función de probabilidad acumulada $F(t)$:

1. Escoger precisión o resolución u con la cual se va a trabajar.
2. Establecer i_{max} como el entero siguiente a $-\log_2(u/rango(t))$, donde $rango(t)$ es el rango posible de t simulados.
3. Inicializar t a 0 para todas las observaciones.
4. Generar 1 número aleatorio U entre 0 y 1 para cada observación
5. Inicializar $i = 1$
6. Establecer $t = t + \frac{rango(t)}{2^i}$ si $U \geq F(t)$, $t = t - \frac{rango(t)}{2^i}$ caso contrario, donde $F(t)$ es la función de probabilidad acumulada de t
7. Establecer $i = i + 1$
8. Repetir desde el paso 6 hasta $i = i_{max}$

Este método converge a la distribución deseada siempre y cuando se tenga un dominio finito.

7.3. Simulador de Datos

```
library(data.table)
library(RcppTN)
library(splines2)
library(stringr)

simular_data = function(n,
                        seed = NULL,
                        ubound = 10,
                        beta_h = 2,
                        beta_p = c(2, 0.5),
                        gamma = c(1, 1.5, 3),
                        knots = NULL,
                        t_prec = 21) {
  cumhaz = function(x, p) {
    return((iSpline(x, knots = knots, intercept = F,
                    Boundary.knots = c(0, ubound * 1.001)) %*%
              gamma) * p)
  }
  sup = function(x, p) {
    return(exp(-cumhaz(x, p)))
  }
  if (!is.null(seed)) {
    set.seed(seed)
  }
  sim = data.table(id = 1:n)
  for (j in 1:length(beta_h)) {
    set(x = sim, j = str_c('xh_', j),
        value = rtn(.mean = rep(0, n),
                    .low = rep(-1.5, n),
                    .high = rep(1.5, n)))
  }
  for (j in 1:length(beta_p[-1])) {
    set(x = sim, j = str_c('xp_', j),
        value = rtn(.mean = rep(0, n),
                    .low = rep(-1.5, n),
                    .high = rep(1.5, n)))
  }
  sim[, e_xb :=
        exp(as.matrix(subset(sim,
                             select = str_subset(names(sim),
                                                 'xh_'))) %*%
              beta_h)]
  sim[, xb_p :=
        as.matrix(subset(sim,
                         select = str_subset(names(sim),
                                             'xp_'))) %*%
```

```
        beta_p[-1] + beta_p[1]]
  sim[, F_t := runif(.N)]
  sim[, p0 := plogis(xb_p)]
  sim[, t := 0]
  for (i in 1:t_prec) {
    sim[, t := t + ubound / (2 ** i) *
          ifelse(F_t < (1 - sup(t, e_xb)) * p0, -1, 1)]
  }
  sim[, t := round(t, 5)]
  sim[, y := runif(.N, 0, ubound)]
  sim[, event := as.numeric(t < y)]
  sim[, time := pmin(t, y)]
  return(list(data = sim,
              bh = beta_h,
              bp = beta_p,
              g = gamma))
}
```

7.4. Simulador de Datos con Newton-Raphson

```
library(data.table)
library(RcppTN)
library(splines2)
library(stringr)

simular_dataNR = function(n,
                          seed = NULL,
                          ubound = 10,
                          beta_h = 2,
                          beta_p = c(2, 0.5),
                          gamma = c(1, 1.5, 3),
                          knots = NULL) {
  iter = function(x, F_x, e_xb) {
    isp = iSpline(x, knots = knots, intercept = F,
                  Boundary.knots = c(0, ubound * 1.001))
    msp = iSpline(x, knots = knots, intercept = F,
                  Boundary.knots = c(0, ubound * 1.001), derivs = 1)
    return(x - (isp %*% gamma + log(1-F_x) / e_xb) / (msp %*% gamma))
  }
  if (!is.null(seed)) {
    set.seed(seed)
  }
  sim = data.table(id = 1:n)
  for (j in 1:length(beta_h)) {
    set(x = sim, j = str_c('xh_', j),
        value = rtn(.mean = rep(0, n),
                    .low = rep(-1.5, n),
                    .high = rep(1.5, n)))
  }
  for (j in 1:length(beta_p[-1])) {
    set(x = sim, j = str_c('xp_', j),
        value = rtn(.mean = rep(0, n),
                    .low = rep(-1.5, n),
                    .high = rep(1.5, n)))
  }
  sim[, e_xb :=
        exp(as.matrix(subset(sim,
                             select = str_subset(names(sim),
                                                 'xh_'))) %*%
              beta_h)]
  sim[, xb_p :=
        as.matrix(subset(sim,
                         select = str_subset(names(sim),
                                             'xp_'))) %*%
        beta_p[-1] + beta_p[1]]
  sim[, F_t := runif(.N)]
```

```
  sim[, p0 := plogis(xb_p)]
  sim[, t := ubound / 2]
  sim[, locked := FALSE]
  repeat {
    sim[!(locked), new_t :=
          pmax(pmin(iter(t, pmin(F_t / p0, 1), e_xb), ubound), 0)]
    sim[!(locked), locked := abs(t - new_t) < 1e-6]
    if (all(sim$locked)) break
    sim[, t := new_t]
  }
  sim[, t := round(t, 5)]
  sim[, y := runif(.N, 0, ubound)]
  sim[, event := as.numeric(t < y)]
  sim[, time := pmin(t, y)]
  return(list(data = sim,
              bh = beta_h,
              bp = beta_p,
              g = gamma))
}
```

7.5. Función de log-verosimilud negativa

```
library(stringr)
library(splines2)

spph_loglik = function(pars,
                       data,
                       n_g = 6,
                       n_n = 3,
                       n_b = 4,
                       knots = c(2, 4, 6)) {
  g = pars[1:n_g]
  d = data$event
  n = pars[(n_g + 1):(n_g + n_n)]
  b = pars[(n_g + n_n + 1):(n_g + n_n + n_b)]
  ubound = max(data$time)
  phi = exp(as.matrix(subset(x = data,
                             select = str_subset(names(data),
                                                 'xh_'))) %*% b)
  z_n = as.matrix(subset(x = data,
                         select = str_subset(names(data),
                                             'xp_'))) %*%
        n[-1] + n[1]
  isp = iSpline(data$time,
                knots = knots,
                Boundary.knots = c(0, ubound))
  msp = iSpline(data$time,
                knots = knots,
                Boundary.knots = c(0, ubound),
                derivs = 1)
  return(-sum(d*(-(isp %*% g) * phi) -
         d*log(1 + exp(-z_n)) +
         d * log((msp %*% g) * phi) +
         (1 - d) * log(1 + (exp(-(isp %*% g) * phi) - 1) /
         (1 + exp(-z_n)))))
}
```

7.6. Algoritmo EM sobre datos simulados

```
source('simular_data.R')
source('spph_loglik.R')
sim_lst = simular_data(1e4, seed = 12345)
Xh = model.matrix(as.formula(str_c('time ~ 0 + ',
                                   str_c(str_subset(names(sim_lst$data),
                                                    'xh_'),
                                         collapse = ' + '))),
                  sim_lst$data)
Xp = model.matrix(as.formula(str_c('time ~ 1 + ',
                                   str_c(str_subset(names(sim_lst$data),
                                                    'xp_'),
                                         collapse = ' + '))),
                  sim_lst$data)
isp = iSpline(sim_lst$data$time,
              knots = c(2, 4, 6),
              Boundary.knots = c(0, 10))
msp = iSpline(sim_lst$data$time,
              knots = c(2, 4, 6),
              Boundary.knots = c(0, 10),
              derivs = 1)
beta_p = rep(0.5, length(sim_lst$bp))
beta_h = rep(0.5, length(sim_lst$bh))
gamma = rep(0.5, length(sim_lst$g))
steps = matrix(data = c(gamma, beta_p, beta_h), nrow = 1)
Rcpp::sourceCpp('pnll.cpp')
i = 1
llik = -Inf
repeat {
  pi_ = 1 - plogis(Xp %*% beta_p)
  pz_1 = pi_ / (pi_ + (1 - pi_) *
         exp(-isp %*% gamma * exp(Xh %*% beta_h)))
  beta_p = -RcppNumerical::fastLR(x = Xp,
                                  y = pz_1 *
                                      (1 - sim_lst$data$event),
                                  start = beta_p)$coefficients
  bg = part_LogLik(Xh,
                   isp,
                   msp,
                   pz_1,
                   sim_lst$data$event,
                   c(gamma, beta_h))
  beta_h = tail(bg, length(beta_h))
  gamma = head(bg, length(gamma))
  steps = rbind(steps, c(gamma, beta_p, beta_h))
  llik_n = -spph_loglik(steps[i + 1, ], data = sim_lst$data)
  if (1 - llik_n / llik < 1e-6 | i == 100) {
```

```
    break
  }
  i = i + 1
  llik = llik_n
}
rm(llik_n)
steps_dt = data.table(steps)
names(steps_dt) = c(paste0('gamma', 1:6),
                    paste0('eta', 0:2),
                    paste0('beta', 1:4))
steps_dt[, Iteracion := 1:.N]
steps_dt = melt(steps_dt,
                id.vars = 'Iteracion',
                measure.vars = names(steps_dt)[-14],
                variable.name = 'Parametro',
                value.name = 'Valor')
ggplot2::ggplot(steps_dt,
                ggplot2::aes(Iteracion,
                             Valor,
                             col = Parametro,
                             group = Parametro)) +
ggplot2::geom_line(lwd = 1.2) +
ggplot2::theme_minimal()
```

7.7. Función de log-verosimilud negativa parcial

```
#include <RcppNumerical.h>
using namespace Numer;
typedef Eigen::Map<Eigen::MatrixXd> MapMat;
typedef Eigen::Map<Eigen::VectorXd> MapVec;
class partLogLik: public MFuncGrad
{
private:
  const MapMat X;
  const MapMat ISP;
  const MapMat MSP;
  const MapVec PZ1;
  const MapVec DEL;
public:
  partLogLik(const MapMat x_,
             const MapMat isp_,
             const MapMat msp_,
             const MapVec pz1_,
             const MapVec del_) : X(x_),
                                  ISP(isp_),
                                  MSP(msp_),
                                  PZ1(pz1_),
                                  DEL(del_) {}
  double f_grad(Constvec& pars, Refvec grad)
  {
    Eigen::VectorXd beta = pars.tail(X.cols());
    Eigen::VectorXd gamma = pars.head(ISP.cols());
    Eigen::VectorXd xbeta = X * beta;
    for(int i = 0; i < 6; i++) {
      if(gamma(i) < 1.0e-10) {
        gamma(i) = 1.0e-10;
      }
    }
    Eigen::VectorXd mgamm = MSP * gamma;
    Eigen::VectorXd igamm = ISP * gamma;
    const double f = -(igamm.array() *
                       xbeta.array().exp() *
                       (PZ1.array() * (1 - DEL.array()) - 1) +
                       DEL.array() *
                       (xbeta.array() +
                        mgamm.array().log())).sum();
    // Gradient
    Eigen::VectorXd grad_val(10);
    for(int i = 0; i < 6; ++i) {
      grad_val(i) = -(ISP.col(i).array() *
                      xbeta.array().exp() *
                      (PZ1.array() * (1 - DEL.array()) - 1) +
```

```
                        DEL.array() *
                        MSP.col(i).array() /
                        mgamm.array()).sum();
    }
    for(int i = 0; i < 4; ++i) {
      grad_val(i + 6) = -(igamm.array() *
                          X.col(i).array() *
                          xbeta.array().exp() *
                          (PZ1.array() * (1 - DEL.array()) - 1) +
                          DEL.array() *
                          X.col(i).array()).sum();
    }
    grad.noalias() = grad_val;
    return f;
  }
};

// [[Rcpp::export]]
Rcpp::NumericVector part_LogLik(Rcpp::NumericMatrix x,
                                Rcpp::NumericMatrix isp,
                                Rcpp::NumericMatrix msp,
                                Rcpp::NumericVector pz1,
                                Rcpp::NumericVector del,
                                Rcpp::NumericVector init)
{
  const MapMat X_ = Rcpp::as<MapMat>(x);
  const MapMat ISP_ = Rcpp::as<MapMat>(isp);
  const MapMat MSP_ = Rcpp::as<MapMat>(msp);
  const MapVec PZ1_ = Rcpp::as<MapVec>(pz1);
  const MapVec DEL_ = Rcpp::as<MapVec>(del);
  // Negative log likelihood
  partLogLik nll(X_, ISP_, MSP_, PZ1_, DEL_);
  // Initial guess
  Eigen::VectorXd parss(init.size());
  for(int i = 0; i < init.size(); i++) {
    parss(i) = init(i);
  }
  double fopt;
  int status = optim_lbfgs(nll, parss, fopt);
  if(status < 0)
    Rcpp::stop("fail to converge");
  return Rcpp::wrap(parss);
}
```

7.8. Maximización de verosimilitud de datos simulados

```
source('simular_data.R')
source('spph_loglik.R')
sim_lst = simular_data(1e4, seed = 12345)
init_pars = rep(0.5, 13)
opt_pars = nlminb(init_pars,
                  spph_loglik,
                  data = sim_lst$data,
                  control = list(trace = 1))
```

7.9. Simular 500 bases de datos

```
cl = parallel::makeCluster(parallel::detectCores())
doSNOW::registerDoSNOW(cl)
N = 500
init_pars = rep(0.5, 13)
pb = txtProgressBar(max = N, style = 3)
progress = function(n) setTxtProgressBar(pb, n)
set.seed(12345)
result = foreach::foreach(i = 1:N,
                          .options.snow = list(progress = progress),
                          .packages = c('stringr',
                                        'splines2',
                                        'data.table',
                                        'RcppTN'),
                          .export = c('spph_loglik',
                                      'simular_data',
                                      'init_pars')) %dopar% {
  sim_lst = simular_data(1e4)
  opt_pars = nlminb(init_pars, spph_loglik, data = sim_lst$data)
  return(list(par = opt_pars$par, counts = opt_pars$counts))
}
close(pb)
stopCluster(cl)
mat = matrix(unlist(lapply(result,
                           function(x) x[['par']])),
             byrow = T,
             nrow = N)
apply(mat, 2, mean)
colnames(mat) = c(str_c('g_', 1:6),
                  str_c('n_', 1:3),
                  str_c('b_', 1:4))
mat = mat[, order(colnames(mat))]
ggplot2::ggplot(melt(as.data.table(mat),
                     measure.vars = colnames(mat)),
                ggplot2::aes(variable,
                             value,
                             col = variable)) +
```

```
ggplot2::geom_point(position = ggplot2::position_jitter(),
                    alpha = 0.2) +
ggplot2::coord_flip()
```

Bibliografía

Joseph Berkson and Robert P Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

John W Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53, 1949.

Vincent Bremhorst and Philippe Lambert. Flexible estimation in cure survival models using bayesian p-splines. *Computational Statistics & Data Analysis*, 93:270–284, 2016.

Li Cai. Sem of another flavour: Two new applications of the supplemented em algorithm. *British Journal of Mathematical and Statistical Psychology*, 61(2):309–329, 2008.

Li Cai and Taehun Lee. Covariance structure model fit testing under missing data: An application of the supplemented em algorithm. *Multivariate Behavioral Research*, 44(2):281–304, 2009.

Vinicius F Calsavara, Agatha S Rodrigues, Ricardo Rocha, Vera Tomazella, and Francisco Louzada. Defective regression models for cure rate modeling with interval-censored data. *Biometrical Journal*, 2019.

Chin Long Chiang. *The life table and its applications*. Krieger Malabar, FL, 1984.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Haskell B Curry and Isaac J Schoenberg. On polya frequency functions iv: The fundamental spline functions and their limits. Technical report, WISCONSIN UNIV MADISON MATHEMATICS RESEARCH CENTER, 1965.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

Vernon T Farewell. A model for a binary variable with time-censored observations. *Biometrika*, 64 (1):43–46, 1977.

Guillaume Gerber, Yohann Le Faou, Olivier Lopez, and Michael Trupin. The impact of churn on client value in health insurance, evaluation using a random forest under random censoring. 2018.

Frank E. Harrell. *Cox Proportional Hazards Regression Model*, pages 465–507. Springer New York, New York, NY, 2001. ISBN 978-1-4757-3462-1. doi: 10.1007/978-1-4757-3462-1_19. URL https://doi.org/10.1007/978-1-4757-3462-1_19.

Tina Harrison and Jake Ansell. Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities. *Journal of Financial Services Marketing*, 6(3):229–239, 2002.

JL Haybittle. A two-parameter model for the survival curve of treated cancer patients. *Journal of the American Statistical Association*, 60(309):16–26, 1965.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Bart Larivière and Dirk Van den Poel. Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, 27(2):277–285, 2004.

Katharina Morik and Hanna Köpcke. Analysing customer churn in insurance data–a case study. In *European conference on principles of data mining and knowledge discovery*, pages 325–336. Springer, 2004.

RF Mould and JW Boag. A test of several parametric statistical models for estimating success rate in the treatment of carcinoma cervix uteri. *British journal of cancer*, 32(5):529, 1975.

T Orchard and MA Woodbury. Missing information principle: Theory and applications. in proceedings of the 6th berke ley symposium on mathematical statistics and probability, theory of statistics, 1972.

Edwin MM Ortega, Vicente G Cancho, and Gilberto A Paula. Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, 15(1):79, 2009.

Donald A Pierce, William H Stewart, and Kenneth J Kopecky. Distribution-free regression analysis of grouped survival data. *Biometrics*, pages 785–793, 1979.

James O Ramsay et al. Monotone regression splines in action. *Statistical science*, 3(4):425–441, 1988.

Sahar F Sabbeh. Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications*, 9(2), 2018.

Pao-sheng Shen, Hsin-Jen Chen, Wen-Harn Pan, and Chyong-Mei Chen. Semiparametric regression analysis for left-truncated and interval-censored data without or with a cure fraction. *Computational Statistics & Data Analysis*, 2019.

Shaun Wang. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*, 17(1):43–54, 1995.

Edward J Wegman and Ian W Wright. Splines in statistics. *Journal of the American Statistical Association*, 78(382):351–365, 1983.

Ken Kwong-Kay Wong. Using cox regression to model customer time to churn in the wireless telecommunications industry. *Journal of Targeting, Measurement and Analysis for marketing*, 19(1): 37–43, 2011.