PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE GRADUADOS

UNA APLICACION DE INTERVALOS DE CONFIANZA PARA LA MEDIANA DE SUPERVIVENCIA EN EL MODELO DE REGRESION DE COX

TESIS PARA OPTAR POR EL GRADO DE MAGISTER EN ESTADÍSTICA

Presentado por:

Jorge Adolfo Mondragón Arboccó

Asesora: Elizabeth Doig Camino

Miembros del jurado:
Dr. Luis Valdivieso Serrano
Dr. Christian Bayes Rodríguez

Lima, Septiembre 2013

Dedicatoria

A mi primo Víctor Andrés con el mejor de los recuerdos.

Agradecimientos

Un sincero y especial agradecimiento a la profesora Elizabeth, sin la cual no hubiera podido concluir este estudio.

Resumen

El presente trabajo estudiará el método propuesto por Tze y Zheng (2006) aplicándolo a la obtención de intervalos de confianza para la mediana de supervivencia de líneas móviles de una empresa de telecomunicaciones. Esta metodología se aplicará con el objeto de conocer el riesgo de vida promedio de la línea móvil así como de qué manera inciden las covariables sobre el tiempo hasta el incumplimiento del pago de los clientes de la empresa.

Para ello se hará uso de una extensión del modelo de Cox haciendo uso de la estimación máximo verosímil para obtener nuevas estimaciones del vector de parámetros mediante el método bootstrap lo que permita la construcción de los intervalos de confianza para la mediana de supervivencia.

Palabras-clave: Modelo de riegos proporcionales, Mediana de supervivencia, Intervalos de confianza basados en pruebas estadísticas, Bootstrap.

Abstract

This paper studies the method proposed by Tze y Zheng (2006) applying it to obtaining confidence intervals for the median survival of mobile lines of a telecommunications company. This methodology is applied in order to know the average risk of the phone line so how covariates influence the time to failure to pay the company's customers.

For this will work with an extention of Cox model using maximum likelihood estimation for get new estimations of the parameters vector through the bootstrap method that will permit the building of intervals confidence for the survival median.

Keywords: Proportional hazards model, Median survival, Test-based confidence intervals, Bootstrap.

Índice general

Lista de Abreviaturas

fdp	función de densidad de probabilidad.
i.i.d.	independientes e idénticamente distribuidas.
c	conjunto de observaciones censuradas.
nc	conjunto de observaciones no censuradas.
PL	estimador de producto límite o de Kaplan-Meier.
v.a.	variable aleatoria.

Lista de Símbolos

$\boldsymbol{\beta}$	vector de parámetros.
$\boldsymbol{x}$	vector de covariables.
$\boldsymbol{Y}$	variable que denota el tiempo de supervivencia.
$\boldsymbol{Y_c}$	tiempo fijo preasignado.
$\boldsymbol{Y_{(i)}}$	estadísticos de orden.
$\boldsymbol{C_i}$	tiempo de censura asociado al individuo i.
$\boldsymbol{\delta_i}$	variables que recogen la información de censura del individuo i.
$\boldsymbol{i}$	información de Fisher.
$\xrightarrow{a}$	convergencia asintótica.
$\xrightarrow{D}$	convergencia en distribución o débil.
$\xrightarrow{c.s.}$	convergencia fuerte o casi segura.
$\xrightarrow{P}$	convergencia en probabilidad.
$\boldsymbol{\Omega}$	conjunto de resultados de un experimento aleatorio.
$\digamma$	es una σ-álgebra de Ω.
$\wedge$	mínimo entre las variables que se comparan.
$\approx$	aproximadamente distribuido.
$\overset{a}{\sim}$	asintóticamente distribuido.
$\boldsymbol{\xi_p(x)}$	p-ésimo cuantil dado el vector de covariables.
$\boldsymbol{c_\alpha\,(t)}$	cuantil bootstrap.
$\boldsymbol{l\,(\beta)}$	función de log-verosimilitud.
$\ddot{\boldsymbol{l}}$	denota la segunda derivada.
$\boldsymbol{\Lambda\,(t)}$	función de riesgo base o subyacente.
$\boldsymbol{\Lambda\,(t \mid x)}$	función de riesgo acumulada asociada al vector de covariables.
$\boldsymbol{\lambda\,(t)}$	tasa de riesgo.
$\boldsymbol{S\,(t \mid x)}$	función de supervivencia asociada al vector de covariables.

Índice de figuras

Capítulo 1

Introducción

Para estudiar los tiempos de supervivencia de un grupo de pacientes frecuentemente se usa la mediana y los intervalos de confianza asociados. Es importante mencionar que son pocos los trabajos que han considerado la presencia de covariables cuando los pacientes provienen de una población homogénea, como es el caso del modelo de riesgos proporcionales.
Hoy en día con los constantes avances de la tecnología, las empresas del rubro de las telecomunicaciones necesitan responder de forma rápida a las necesidades de sus clientes. En un mercado como el peruano, en donde la cobertura de la telefonía móvil ha llegado casi a su totalidad (80 de cada 100 habitantes disponen de una línea móvil para comunicarse [1]), las operadoras de telecomunicaciones tienen como tarea principal generar en el mercado las condiciones que muestren una percepción de un servicio de calidad con la finalidad de atraer a los clientes de la competencia y el de conservar a sus clientes actuales. Así pues, el prolongar el tiempo de vida de los clientes constituye una necesidad básica actualmente.
El presente trabajo estudiará el método propuesto por Tze y Zheng (2006) aplicándolo a la obtención de intervalos de confianza para la mediana de supervivencia de líneas móviles de una empresa de telecomunicaciones. Esta metodología se aplicará con el objeto de conocer el riesgo de vida promedio de la línea móvil así de cómo las covariables inciden sobre el tiempo hasta el incumplimiento del pago de los clientes de la empresa.

1.1. Antecedentes

El modelo de riesgos proporcionales de Cox (1972) es un modelo de regresión log-lineal que relaciona a la función de riesgo acumulada $\Lambda(t \mid \boldsymbol{x})$ con un conjunto de covariables de la siguiente manera:

$$\Lambda(t \mid \boldsymbol{x}) = \Lambda(t) \exp\left(\beta' \boldsymbol{x}\right) \tag{1.1}$$

donde:
$\boldsymbol{x}$ es el vector de covariables
β es el vector de parámetros asociado a $\boldsymbol{x}$
$\Lambda(t)$ es la función de riesgo subyacente o base

Basándose en una muestra que consta de n observaciones $(\widetilde{t}_i, \delta_i, x_i)$, donde $\widetilde{t}_i = \text{mín}(t_i, c_i)$ y

[1] Fuente: Organismo Supervisor de la Inversión Privada en Telecomunicaciones (2012)

$\delta_i = I_{\{t_i \leq c_i\}}$ es el indicador de si el actual tiempo de falla t_i es observado o censurado (c_i), se obtiene la estimación $\widehat{\beta}$ de β que maximiza la función de log-verosimilitud parcial:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta' x_i - \log \left(\sum_{j:\widetilde{t}_j \geq \widetilde{t}_i} \exp\left(\beta' x_j\right) \right) \right\}. \tag{1.2}$$

Regiones de confianza para β pueden ser construidas usando la normalidad asintótica de $\left(-\ddot{l}\left(\widehat{\beta}\right)\right)^{-1/2}\left(\widehat{\beta}-\beta\right)$ o la distribución χ^2 límite de $2\left\{l\left(\widehat{\beta}\right) - l(\beta)\right\}$, donde $\ddot{l}$ denota la segunda derivada. En muchas aplicaciones, es útil estimar también la mediana de los tiempos de supervivencia dado el vector de covariables de los sujetos. En particular, mediante la combinación de $\widehat{\beta}$ con la estimación de la función de riesgo acumulada subyacente de Breslow (1974) $(\widehat{\Lambda})$, Miller y Halpern (1982) usaron la mediana de la función de distribución $1 - \exp\left\{-\widehat{\Lambda}(.)\, e^{\widehat{\beta}' \boldsymbol{x}}\right\}$ para estimar la mediana de supervivencia. Dabrowska y Doksum (1987) y posteriormente Burr y Doss (1993) estudiaron el problema de la construcción de intervalos de confianza para la mediana de los tiempos de supervivencia dadas las covariables de los sujetos. Sea $\xi_p(\boldsymbol{x})$ que denota el p-ésimo cuantil de la distribución del tiempo de falla para un vector de covariables $\boldsymbol{x}$ (entonces $p = 1/2$ corresponde al de la mediana) y $\widehat{\xi}_p(\boldsymbol{x})$ el p-ésimo cuantil estimado de la función de distribución precedente, su enfoque se basa en la aproximación de normalidad de $\left\{\widehat{\xi}_p(\boldsymbol{x}) - \xi_p(\boldsymbol{x})\right\}/\widehat{se}_p(\boldsymbol{x})$, o en su proceso gaussiano límite indexado por $\boldsymbol{x}$, donde $\widehat{se}_p(\boldsymbol{x})$ denota la estimación del error estándar de $\widehat{\xi}_p(\boldsymbol{x})$.

La varianza de la distribución normal límite de $\sqrt{n}\left\{\widehat{\xi}_p(\boldsymbol{x}) - \xi_p(\boldsymbol{x})\right\}$ depende de la función de riesgo subyacente $\lambda(t) = (d/dt)\,\Lambda(t)$, lo que genera una dificultad para los cálculos. Aunque Dabrowska y Doksum (1987) citan a Tsiatis (1981) y Andersen y Gill (1982) alegando consistencia sobre su estimador propuesto de varianza límite, Tsiatis, Anderson y Gill únicamente han establecido consistencia para el estimador de Breslow de Λ pero no de la derivada λ. Burr y Doss (1993) hacen uso de la suavización de $\widehat{\Lambda}$ para estimar λ, y en lugar de aplicar la teoría asintótica de $\left\{\widehat{\xi}_p(\boldsymbol{x}) - \xi_p(\boldsymbol{x})\right\}/\widehat{se}_p(\boldsymbol{x})$ directamente para construir intervalos de confianza para $\xi_p(\boldsymbol{x})$, la usan para proveer una justificación teórica del método bootstrap-t para construir intervalos de confianza. Sin embargo, como ha sido señalado por Efron y Tibshirani (1993), el método bootstrap-t requiere estimaciones estables de los errores estándar para que funcione bien en la práctica. Por tanto, las dificultades en estimar el error estándar de $\widehat{\xi}_p(\boldsymbol{x})$ también causan dificultades con intervalos de confianza bootstrap-t para $\xi_p(\boldsymbol{x})$.

En efecto, incluso sin censura y sin covariables de modo que el problema se reduzca a intervalos de confianza para el p-ésimo cuantil ξ_p de una función de distribución basada en una muestra de tiempos de supervivencia independientes e idénticamente distribuidos $t_1, ..., t_n$ con función de densidad común f que tiene un estimador consistente $\widehat{f}$, la distribución normal limite de

$$\widehat{f}(\xi_p)\left\{n/\left[p(1-p)\right]\right\}^{1/2}\left(\widehat{\xi}_p - \xi_p\right) \tag{1.3}$$

raramente es usada en la construcción de intervalos de confianza para ξ_p. Además de los problemas con el rendimiento de muestras finitas del estimador de densidad $\widehat{f}\left(\widehat{\xi}_p\right)$, la adecuación de la aproximación lineal $f(\xi_p)\left(\widehat{\xi}_p - \xi_p\right)$ a $F\left(\widehat{\xi}_p\right) - F(\xi_p)$ usada para derivar la

normalidad asintótica de $\widehat{\xi}_p - \xi_p$ (donde F es la función de distribución cuya derivada es f) es problemática cuando $\widehat{\xi}_p$ no es suficientemente cercana a ξ_p. En su lugar, un intervalo de confianza estándar no paramétrico de la forma $t_{(k_1)} < \xi_p < t_{(k_2)}$, donde $t_{(i)}$ denota el estadístico de orden de la muestra y $k_1 < k_2$ son enteros tales que:

$$P\left(t_{(k_1)} \leq \xi_p < t_{(k_2)}\right) = P\left(k_1 \leq B\left(n,p\right) < k_2\right) \geq 1 - 2\alpha. \tag{1.4}$$

El límite inferior $1-2\alpha$ en (1.4) puede no ser alcanzable debido a la discontinuidad de la distribución Binomial $B\left(n,p\right)$. Como fue mostrado por Efron (1979) y Chen y Hall (1993), los intervalos de confianza para los percentiles bootstrap y los intervalos de confianza de verosimilitud empírica (obtenidos mediante la inversión de las pruebas de razón de verosimilitud empírica) para ξ_p son de esta forma (ver también Efron y Tibshirani (1993)). Chen y Hall (1993) también muestran que la incapacidad de (1.4) para alcanzar $1-2\alpha$ con un error $O\left(n^{-1}\right)$ debido a la discontinuidad de la distribución Binomial se puede superar mediante el uso de una versión suavizada de verosimilitud empírica. Un método alternativo para lograr una probabilidad de cobertura de $1-2\alpha+O\left(n^{-1}\right)$ fue propuesto por Beran y Hall (1993) quienes utilizaron combinaciones convexas de cuantiles muestrales para desarrollar intervalos de confianza interpolados. Posteriormente Ho y Lee (2005) hicieron uso de iteraciones bootstrap suavizadas para lograr errores de cobertura más precisos de un solo lado del intervalo percentil bootstrap. Este método, sin embargo, es muy intensivo computacionalmente e implica una capa adicional de bootstrapping para determinar el parámetro utilizado para suavizar la distribución empírica.

Para los datos de supervivencia censurados sin covariables, Li et al. (1996) hizo uso de la verosimilitud empírica para construir bandas de confianza para ξ_p, de forma conjunta en $p_1 \leq p \leq p_2$. Sus resultados sobre las probabilidades de cobertura se basan en la convergencia débil y no proporcionan tasas de convergencia de los tipos dados en Chen y Hall (1993). Ellos sin embargo, no suavizaron la función de verosimilitud empírica, ni compararon el enfoque de verosimilitud empírica con otros métodos de prueba basados en construir intervalos de confianza para ξ_p cuando los t_i están sujetos a censura. Estos intervalos basados en pruebas estadísticas alternativas se remontan a Brookmeyer y Crowley (1982) quienes invierten una prueba de signos generalizados, que conducen a un conjunto de confianza $1-2\alpha$ aproximado de la forma

$$\left\{t : \left|\widehat{S}\left(t\right) - 1/2\right| \leq z_{(1-\alpha)}\widehat{\sigma}\left(t\right)\right\} \tag{1.5}$$

para la mediana $\xi_{1/2}$, donde $\widehat{S}\left(t\right)$ es el estimador de Kaplan-Meier de la función de supervivencia, $\widehat{\sigma}\left(t\right)$ es el error estándar estimado de $\widehat{S}\left(t\right)$ y z_q denota el q-ésimo cuantil de la distribución normal estándar. En lugar de utilizar la aproximación normal, Strawderman et al. (1997) utiliza expansiones Edgeworth para la función de riesgo acumulada studentizada para obtener límites de confianza basados en pruebas estadísticas más precisas para ξ_p.

En este trabajo se desarrolla un nuevo método para construir intervalos de confianza para el cuantil $\xi_p\left(\boldsymbol{x}\right)$ en el modelo de riesgos proporcionales (1.1). A diferencia de los métodos de Dabrowska y Doksum (1987) y Burr y Doss (1993) que usan $\left\{\widehat{\xi}_p\left(\boldsymbol{x}\right) - \xi_p\left(\boldsymbol{x}\right)\right\}/\widehat{se}_p$ como un pivote aproximado, aquí se usa un enfoque basado en pruebas estadísticas, empleando

$\widehat{\Lambda}(t \mid \boldsymbol{x})$ para probar si $\Lambda(t \mid \boldsymbol{x}) = \log\left(p^{-1}\right)$, donde $\widehat{\Lambda}(t \mid \boldsymbol{x}) = \widehat{\Lambda}(t) \exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{x}\right)$ y donde $\widehat{\Lambda}(t)$ es el estimador de Breslow de la función de riesgo acumulada subyacente $\Lambda(t)$. En lugar de usar la aproximación normal como en Strawderman et al. (1997) para encontrar los cuantiles de la estadística de prueba, se usa el método bootstrap para evaluar los cuantiles de un pivot aproximado obtenido por studentizar la estadística de prueba.

1.2. Objetivos

El objetivo general de la tesis es estudiar e implementar los intervalos de confianza para la mediana de supervivencia en el modelo de riesgos proporcionales de Cox aplicado a un conjunto de datos de una empresa de telecomunicaciones, haciendo uso para ello del método propuesto por Tze y Zheng (2006).
Los objetivos específicos de la tesis son los siguientes:

- Revisar la literatura de los siguientes conceptos teóricos: Análisis de supervivencia, Pruebas estadísticas basadas en intervalos de confianza, Método Bootstrap y Teoría Asintótica.
- Estudiar y profundizar en los conceptos teóricos en los que se basa el modelo propuesto.
- Estudiar e implementar computacionalmente el modelo propuesto.
- Realizar la aplicación del modelo a un conjunto de datos de una empresa de telecomunicaciones.
- Analizar los resultados obtenidos de la aplicación del modelo.

1.3. Organización del trabajo

En el Capítulo 2, se presentan los conceptos referentes a los temas que permiten estudiar el modelo propuesto.
En el Capítulo 3 se revisan las pruebas estadísticas basadas en intervalos de confianza que son las que se emplean para la implementación del modelo, el cuál se detalla en el Capítulo 4 en donde se presentan las ecuaciones y diagramas de flujo asociados a tales pruebas.
En el Capítulo 5 se realiza la aplicación del modelo al conjunto de datos reales de una empresa de telecomunicaciones.
Finalmente, en el Capítulo 6 se discuten las conclusiones obtenidas así como se presentan las ideas y sugerencias para un futuro trabajo de investigación.
En el Apéndice A se presentan las demostraciones y los conceptos teóricos empleados. En el Apéndice B se presentan los programas construidos para el desarrollo del modelo estudiado y su aplicación. Por último, en el apéndice C se muestran algunos de los resultados numéricos obtenidos.

Capítulo 2

Conceptos preliminares

En el presente capítulo se hace un desarrollo de los conceptos que se requieren para estudiar el método que calcula los intervalos de confianza para la mediana de supervivencia que se presentan en el Capítulo 3.

2.1. Análisis de Supervivencia

El análisis de supervivencia estudia los procesos aleatorios relacionados con la muerte de organismos vivos y la falla de sistemas físicos (mecánicos o electrónicos).
Es una rama de la estadística que comprende una variedad de técnicas para estudiar los tiempos de ocurrencia de un evento de interés, como por ejemplo el tiempo de aprendizaje de una habilidad, la detección de una enfermedad, la falla de un equipo, la baja o pérdida de un cliente, la desactivación de una línea, etc.
Éstos métodos lograron un desarrollo a partir del estudio de la confiabilidad de los equipos militares empleando modelos paramétricos (aplicaciones en ingeniería), posteriormente se desarrollaron estudios de investigación con experimentos clínicos, haciéndose uso de modelos no paramétricos.
Para resumir los datos del análisis de supervivencia se tienen dos funciones, la función de supervivencia y la función de riesgo.

2.1.1. Función de Supervivencia y función de Riesgo

Sea la variable aleatoria Y que representa el tiempo hasta que ocurra la muerte de un individuo (por consiguiente su valor es cero o toma valores positivos) cuya función de densidad es $f(y)$ y su función de distribución $F(y)$.
Se define a la función de supervivencia como:

$$\begin{aligned} S(y) &= P[\text{el individuo sobreviva un tiempo mayor al tiempo } y] \\ &= P[Y > y] = 1 - P[Y \leq y] = 1 - F(y) \end{aligned}$$

Y a la función o tasa de riesgo como:

$$\lambda(y) = \frac{f(y)}{1 - F(y)} = \frac{f(y)}{S(y)} \tag{2.1}$$

La cuál se interpreta de la siguiente manera:

$\lambda(y)\,dy \approx P[\text{el individuo expire en el intervalo }]y, y+dy[\ |\ \text{sobrevivió en un tiempo pasado } y]$

Para cada observación es necesario conocer el instante de origen, el instante en que ocurre el evento de interés y el tiempo transcurrido entre éstos, lo que se define como tiempo de supervivencia. Sin embargo, conocer éstos instantes o el tiempo exacto de supervivencia no siempre es posible, produciéndose diferentes tipos de censura en las observaciones.
Para el análisis de supervivencia con observaciones censuradas se han desarrollado procedimientos especiales para tratar los diferentes tipos de censura. Estos modelos constan de una variable que mide el tiempo de supervivencia hasta la ocurrencia del evento de interés o la censura y de otra que indica el tipo de censura que se presenta, se trabaja además bajo el supuesto de que los tiempos de supervivencia y los de censura son independientes.

2.1.2. Tipos de censuramiento

Cuando la información que proporciona una observación sobre el elemento en estudio es incompleta sobre su tiempo de vida (Y) se dice que es censurada.
Generalmente se trabaja con tres tipos de censura:
Sean las variables aleatorias $Y_1, Y_2, ..., Y_n$ independientes e idénticamente distribuidas con función de distribución F.

1. Tipo I de Censuramiento:

   Sea Y_c un tiempo fijo preasignado que se denomina tiempo de censuramiento fijo. Se observan las variables aleatorias:

   $$T_i = \begin{cases} Y_i \text{ , si } Y_i \leq Y_c \\ Y_c \text{ , si } Y_i > Y_c \end{cases} \tag{2.2}$$

   Notar que la $P[T_i = Y_c] = P[Y_i > Y_c] > 0$, ya que $Y \geq 0,\ Y \in [0, \infty+]$. Luego:

   $$T_i = \text{mín}\{Y_i, Y_c\},\ \forall\ i.$$

2. Tipo II de Censuramiento:

   Sea $r < n$ fijo y sean $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$ los estadísticos de orden de $Y_1, Y_2, ..., Y_n$, si en el estudio se observan los tiempos hasta la ocurrencia de la r-ésima muerte, se tiene:

   $$T_{(i)} = \begin{cases} Y_{(i)}, \text{ si } Y_{(i)} \leq Y_{(r)} \\ Y_{(r)}, \text{ si } Y_{(i)} > Y_{(r)} \end{cases} \tag{2.3}$$

   ó

   $$T_{(i)} = \text{mín}\left\{Y_{(i)}, Y_{(r)}\right\}.$$

   Éstos dos tipos de censura se emplean generalmente en aplicaciones a la ingeniería.

3. Censuramiento aleatorio:

   Sean las variables $C_1, C_2, ..., C_n$ independientes e idénticamente distribuidas con función de distribución G, entonces C_i es el tiempo de censura asociado al paciente i con su respectivo tiempo de vida Y_i. Se observa lo siguiente:

$$T_i = \begin{cases} Y_i, \text{ si } Y_i \leq C_i, \\ C_i, \text{ si } Y_i > C_i \end{cases} \tag{2.4}$$

   Se definen las variables:

$$\delta_i = \begin{cases} 1, \text{ si } Y_i \leq C_i, \\ 0, \text{ si } Y_i > C_i \end{cases} \tag{2.5}$$

   Las δ_i recogen la información de la censura, luego la información completa estará dada por $(T_i, \delta_i), \forall i$, además:

$$T_i = \text{mín}\{Y_i, C_i\}, \ \forall i \tag{2.6}$$

   Este tipo de censura se emplea en los ensayos clínicos.

Los tipos más frecuentes en que se presenta la censura son:

1. Perderse de vista, que se refiere a los casos en que el paciente abandona el tratamiento.
2. Retirarse, debido a que el tratamiento tiene consecuencias negativas.
3. Culmina el estudio.

2.1.3. Estimación de los parámetros de los modelos paramétricos por máxima verosimilitud

Se deduce que en el modelo de censuramiento aleatorio (notar que esto incluye el Tipo I de censuramiento al definir $C_i = Y_c$. También, las verosimilitudes para el Tipo II de censuramiento son similares a los del Tipo I de censuramiento excepto para la multiplicación de algunas constantes que toman en cuenta el orden), el par (T_i, δ_i) tiene verosimilitud:

$$\begin{aligned} L(T_i, \delta_i) &= \begin{cases} f(T_i) \ , \text{ si } \delta_i = 1 \ \text{ (no censura)} \\ S(T_i) \ , \text{ si } \delta_i = 0 \ \text{ (censura)} \end{cases} \\ &= f(T_i)^{\delta_i} S(T_i)^{1-\delta_i} \end{aligned}$$

y la verosimilitud de la muestra completa es:

$$L(T_1, ..., T_n; \delta_i, ..., \delta_n) = \prod_{i=1}^{n} L(T_i, \delta_i) = \left(\prod_{nc} f(T_i)\right)\left(\prod_{c} S(T_i)\right) \tag{2.7}$$

donde $\prod_{nc}$ y $\prod_{c}$ denotan el producto de las observaciones no censuradas y censuradas respectivamente. Luego, las verosimilitudes completas bajo censuramiento aleatorio son:

$$L(T_i, \delta_i) = \begin{cases} f(T_i)\,[1 - G(T_i)] & , \text{si } \delta_i = 1 \\ g(T_i)\,S(T_i) & , \text{si } \delta_i = 0 \end{cases}$$

$$L = \left(\prod_{nc} f(T_i)\right)\left(\prod_{c} S(T_i)\right)\left(\prod_{c} g(T_i)\right)\left(\prod_{nc} [1 - G(T_i)]\right)$$

pero bajo el supuesto que el tiempo de censuramiento no tiene conexión con el tiempo de supervivencia, los últimos dos productos $\prod_{c} g(T_i)$ y $\prod_{nc} [1 - G(T_i)]$ no involucran los parámetros de tiempo de vida desconocidos, por lo tanto éstos dos productos pueden ser tratados como constantes cuando se maximiza L.
Sea $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ el vector de parámetros. Encontrar el $\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$ es equivalente a encontrar la solución $\widehat{\boldsymbol{\beta}}$ de las ecuaciones de verosimilitud:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta_j} \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \log L_{\boldsymbol{\beta}}(T_i, \delta_i) \\ &= \sum_{nc} \frac{\partial}{\partial \beta_j} \log f_{\boldsymbol{\beta}}(T_i) + \sum_{c} \frac{\partial}{\partial \beta_j} \log S_{\boldsymbol{\beta}}(T_i)\,, \;\; j = 1, ..., p \end{aligned}$$

Éstas ecuaciones se resuelven empleando el método de Newton-Raphson.

2.1.4. Estimación de funciones

2.1.4.1. El método Actuarial

Podemos descomponer las probabilidades de supervivencia $S(t)$ en el siguiente producto de probabilidades:

$$\begin{aligned} S(t) &= P(T > \tau_k) \\ &= P(T > \tau_1)\, P(T > \tau_2 \mid T > \tau_1) ... P(T > \tau_k \mid T > \tau_{k-1}) \\ &= p_1 p_2 ... p_k \end{aligned}$$

donde: $p_i = P(T > \tau_i \mid T > \tau_{i-1})$.

El método actuarial brinda una estimación para cada p_i de manera separada, luego agrupa las estimaciones y las multiplica para estimar $S(t)$.
Para una estimación de p_i, se puede usar $1 - d_i/n_i$, si no hay pérdidas o retiros en un intervalo dado I_i que son intervalos casi siempre, pero no necesariamente, de la misma longitud, con $I_i =]\tau_{i-1}, \tau_i]$. Sin embargo, con l_i (número de pérdidas durante el intervalo I_i) y w_i (número de retiros durante el intervalo I_i) distintos de cero, se asume que, en promedio, aquellos individuos que se perdieron o retiraron durante I_i están en riesgo durante la mitad del intervalo. Por lo tanto se define el tamaño de la muestra como:

$$n_i^{'} = n_i - \frac{1}{2}\left(l_i + w_i\right),$$

y

$$\widehat{q}_i = d_i/n_i^{'},$$
$$\widehat{p}_i = 1 - \widehat{q}_i$$

Entonces la estimación actuarial es:

$$\widehat{S}\left(t\right) = \prod_{i=1}^{k} \widehat{p}_i. \tag{2.8}$$

Considerar que si una estimación más fina de $S\left(t\right)$ es requerida, el estimador de Producto Límite de Kaplan-Meier es el enfoque a tomar (Miller (1981)).

2.1.4.2. Varianza de $\widehat{S}\left(\tau_k\right)$

Para estimar la varianza de $\widehat{S}\left(\tau_k\right)$ se considera:

$$\log \widehat{S}\left(t\right) = \sum_{i=1}^{k} \log \widehat{p}_i. \tag{2.9}$$

Se asume que $n\widehat{p}_i \approx B\left(n, \widehat{p}_i\right)$, el método delta implica que

$$Var\left(\log \widehat{p}_i\right) \cong Var\left(\widehat{p}_i\right)\left(\frac{d}{dp_i}\left(\log p_i\right)\right)^2 = \frac{p_i q_i}{n}.\frac{1}{p_i^2} = \frac{q_i}{np_i}, \tag{2.10}$$

y asumiendo que los $\log \widehat{p}_i, ..., \log \widehat{p}_k$ son independientes,

$$Var\left[\log \widehat{S}\left(\tau_k\right)\right] = \sum_{i=1}^{k} \frac{q_i}{np_i}, \tag{2.11}$$

$$Var\left[\log \widehat{S}\left(\tau_k\right)\right] = \sum_{i=1}^{k} \frac{\widehat{q}_i}{n\widehat{p}_i} = \sum_{i=1}^{k} \frac{d_i}{n\left(n - d_i\right)}. \tag{2.12}$$

Finalmente usamos el método Delta nuevamente para obtener:

$$Var\left[\widehat{S}\left(\tau_k\right)\right] = \widehat{S}^2\left(\tau_k\right) \sum_{i=1}^{k} \frac{d_i}{n\left(n - d_i\right)}, \tag{2.13}$$

que es llamada la fórmula de Greenwood.

2.1.4.3. Estimador de Producto Límite (Kaplan-Meier)

El estimador de Producto Límite es similar al estimador actuarial excepto que las longitudes de los intervalos son variables.
Sea τ_i el extremo derecho del intervalo I_i (esto coincide con el tiempo del dato censurado o no censurado), se divide el tiempo en una secuencia de intervalos $I_1, ..., I_k$ y se asume que no

hay empates en los datos. Luego, se ordenan los tiempos T_i de menor a mayor, con lo que se tienen los estadísticos de orden $T_{(1)}, .., T_{(n)}$. Los tiempos T_i indican la ocurrencia de o no de la muerte o falla del individuo.

Notar que $\delta_{(i)}$ es el valor de δ asociado a los estadísticos de orden $T_{(i)}$ donde $\delta_{(i)} = \delta_j \; si \; T_{(i)} = T_j$.

También se tiene que:

$n_i =$ número de personas que están en $R\left(T_{(i)}\right)$ es decir que están vivas

$d_i =$ número de personas fallecidas en el tiempo $T_{(i)}$

$p_i = P\left[\text{de que sobrevivan en el intervalo } I_i \mid \text{están vivas al inicio de } I_i\right]$

$\quad = P\left[Y > \tau_i \mid Y > \tau_{i-1}\right]$

$q_i = 1 - p_i$

Los estimadores:

$$\widehat{q}_i = \frac{d_i}{n_i}$$

$$\widehat{p}_i = 1 - \widehat{q}_i = \begin{cases} 1 - \frac{1}{n_i}, \text{ si } \delta_{(i)} = 1 \text{ (no censura)} \\ 1, \text{ si } \delta_{(i)} = 0 \text{ (censura)} \end{cases}$$

Luego: $\widehat{S}(t) = \prod_{\{i|T_{(i)} \le t\}} \widehat{p}_i = \prod_{\{i \in nc | T_{(i)} \le t\}} \left(1 - \frac{1}{n_i}\right)$, si n_i representa al conjunto de personas en riesgo para el instante $T_{(i)}$, donde $T_{(i)}$ es el tiempo de muerte $(n_i = n - (i-1))$ y nc es el conjunto de observaciones no censuradas.

Luego: $\widehat{S}(t) = \prod_{\{i \in nc | T_{(i)} \le t\}} \left(1 - \frac{1}{n-i+1}\right) = \prod_{\{i \in nc | T_{(i)} \le t\}} \left(\frac{n-i}{n-i+1}\right)$, si no hay empates.

Notar que:

1. Para el caso de las observaciones no censuradas con empates: Se puede suponer que antes del tiempo t hay "m" individuos vivos y en el tiempo "t" ocurren "d" muertes, entonces se divide al tiempo de las muertes de manera infinitesimal de tal forma que el factor para las "d" muertes en el estimador de Kaplan-Meier es de la forma:

$$\left(1 - \frac{1}{m}\right)\left(1 - \frac{1}{m-1}\right)\left(1 - \frac{1}{m-2}\right) ... \left(1 - \frac{1}{m-d+1}\right) =$$
$$\left(\frac{m-1}{m}\right)\left(\frac{m-2}{m-1}\right)\left(\frac{m-3}{m-2}\right) ... \left(\frac{m-d+1}{m-d+2}\right)\left(1 - \frac{m-d}{m-d+1}\right) =$$
$$\left(\frac{m-d}{m}\right) =$$
$$1 - \frac{d}{m}$$

2. Si las observaciones censuradas y no censuradas están empatadas, se consideran a las observaciones no censuradas como si ocurrieran antes de las observaciones censuradas.

3. Si la última observación $T_{(n)}$ es censurada, entonces el límite: $\lim_{t\to\infty} \widehat{S}(t) > 0$. Por costumbre conviene redefinir a $\widehat{S}(t) = 0,\ \forall\ T_{(n)} \leq t$, o se asume que $\widehat{S}(t)$ no está definida $\forall\ t \geq T_{(n)}$.

4. En base a los puntos dados $\left\{T_{(i)},\ 1 \leq i \leq n\right\}$, se definen los nuevos valores $T^*_{(1)} < ... < T^*_{(r)}$, que corresponden a los tiempos distintos en el conjunto total de la muestra y donde "r" representa el número total de éstos tiempo distintos.

$$\delta^*_{(j)} = \begin{cases} 1, \text{ si las observaciones en } T^*_{(j)} \text{ son no censuradas} \\ 0, \text{ si las observaciones en } T^*_{(j)} \text{ son censuradas} \end{cases}$$

$n^*_j =$ número de sujetos vivos en riesgo que están en $R\left(T^*_{(j)}\right)$
$d^*_j =$ número de sujetos muertos en el tiempo $T^*_{(j)}$

Luego $\widehat{S}(t) = \prod_{\left\{j \in nc | T^*_{(j)} \leq t\right\}} \left(1 - \frac{d^*_j}{n^*_j}\right)$, es la forma de Kaplan-Meier con empates.

2.1.4.4. Varianza de $\widehat{S}(t)$

De acuerdo a los resultados obtenidos en Miller (1981) se tiene:

$$Var\left[\widehat{S}(t)\right] = \widehat{S}(t)^2 \sum_{\{i|T_{(i)} \leq t\}} \frac{\widehat{q}_i}{n\widehat{p}_i} = \widehat{S}(t)^2 \sum_{\{i|T_{(i)} \leq t\}} \frac{d_{(i)}}{(n-i)(n-i+1)}, \text{ cuando no hay empates} \tag{2.14}$$

$$Var\left[\widehat{S}(t)\right] = \widehat{S}(t)^2 \sum_{\{j|T_{(j)} \leq t\}} \frac{\delta^*_{(j)} d^*_j}{n^*_j\left(n^*_j - d^*_j\right)}, \text{ cuando hay empates} \tag{2.15}$$

Las cuáles son conocidas como las fórmulas de Greenwood.

2.1.4.5. Normalidad Asintótica

Sean F y G funciones continuas en $[0, T]$ y $F(T) > 0$, $F(T) < 1$ (Miller (1981)), entonces:

$$Z_n(t) = \sqrt{n}\left[\widehat{S}(t) - S(t)\right] \xrightarrow{D} \{Z(t)\} \text{ cuando } n \to \infty, \tag{2.16}$$

donde $\{Z(t)\}$ es un proceso gaussiano con momentos:

$$\begin{aligned} E(Z(t)) &= 0, \\ Cov(Z(t_1), Z(t_2)) &= S(t_1) S(t_2) \int_0^{\min\{t_1,t_2\}} \frac{dF_{nc}(u)}{[1-H(u)]^2}, \\ &= S(t_1) S(t_2) \int_0^{\min\{t_1,t_2\}} \frac{dF_{nc}(u)}{[1-F(u)][1-H(u)]}, \end{aligned}$$

donde:

$$F_{nc}(t) = P\{T \leq t, \delta = 1\} = \int_0^t [1 - G(u)]\, dF(u),$$
$$1 - H(u) = [1 - F(u)]\,[1 - G(u)].$$

Notar que cuando $\{Z_n(t)\}$ converge débilmente $\left(\overset{D}{\rightarrow}\right)$ a un proceso gaussiano $\{Z(t)\}$ significa que para cualquier selección $t_1, ..., t_k$, las variables $Z_n(t_1), ..., Z_n(t_k)$ tienen una distribución normal multivariante. Además $f(Z_n)$ converge en distribución a $f(Z)$ para cualquier distribución f continua.
Como un caso particular del resultado anterior,

$$\widehat{S}(t) \overset{a}{\rightarrow} N\left(S(t), \frac{S^2(t)}{n} \int_0^t \frac{dF_{nc}(u)}{[1 - H(u)]^2}\right). \tag{2.17}$$

Si se deseara obtener una aproximación para la varianza asintótica de $\widehat{S}(t)$, se puede partir de lo siguiente:

$$d\widehat{F}_{nc}\left(t_{(i)}\right) = \frac{\delta_{(i)}}{n},$$
$$1 - \widehat{H}\left(t_{(i)}\right) = 1 - \frac{i}{n} = \frac{n-i}{n},$$
$$1 - \widehat{H}\left(t_{(i)^-}\right) = 1 - \frac{i-1}{n} = \frac{n-i+1}{n}.$$

Entonces el estimador de la varianza asintótica será:

$$\widehat{AVar}\left(\widehat{S}(t)\right) = \frac{\widehat{S}^2(t)}{n} \sum_{\{i/t_{(i)} \leq t\}} \frac{\delta_{(i)}/n}{[(n-i)/n]\,[(n-i+1)/n]},$$
$$= \widehat{S}^2(t) \sum_{\{i/t_{(i)} \leq t\}} \frac{\delta_{(i)}}{[(n-i)]\,[(n-i+1)]},$$

ésta fórmula coincide con la fórmula de Greenwood. $\widehat{AVar}$ denota aquí a la varianza asintótica.

2.1.4.6. Estimadores de la función de riesgo

Para estimar la función de riesgo $\lambda(t)$ expresada en la ecuación (2.1) se estimará a la función:

$$\Lambda(t) = \int_0^t \lambda(u)\, du. \tag{2.18}$$

Las funciones Λ y S están relacionadas por: $S(t) = e^{-\Lambda(t)}$.
Considerando que no hay empates en las observaciones, Nelson (1969) estima $\lambda(t)$ por:

$$\widehat{\Lambda}_2(t) = \sum_{\{i/t(i) \leq t\}} \frac{\delta_{(i)}}{n - i + 1}, \tag{2.19}$$

y Peterson (1977) propone:

$$\widehat{\Lambda}_1(t) = \sum_{\{i/t(i)\leq t\}} -\log\left(1 - \frac{\delta_{(i)}}{n-i+1}\right). \tag{2.20}$$

El estimador de Peterson corresponde al estimador Producto Límite de la función de supervivencia

$$\widehat{S}_1(t) = e^{-\widehat{\Lambda}_1(t)} = \prod_{\{i/t(i)\leq t\}} \left(1 - \frac{\delta_{(i)}}{n-i+1}\right), \tag{2.21}$$

mientras que el estimador de Nelson corresponde a un estimador diferente de la función de supervivencia:

$$\widehat{S}_2(t) = e^{-\widehat{\Lambda}_2(t)}. \tag{2.22}$$

Fleming y Harrington (1979) recomiendan $\widehat{S}_2(t)$ como un estimador alternativo para la función de supervivencia y muestran que tiene un error cuadrático medio ligeramente más pequeño en algunas situaciones.

2.1.5. Modelo de regresión de riesgos proporcionales de Cox

Sea $\lambda(t \mid \boldsymbol{x})$ la tasa de riesgo para un individuo en el tiempo t asociado con un conjunto de covariables $\boldsymbol{x}$. El modelo básico dado por Cox (1972) es:

$$\lambda(t \mid \boldsymbol{x}) = \lambda(t)\, c\left(\beta' \boldsymbol{x}\right) \tag{2.23}$$

donde $\lambda(t)$ es la función de riesgo base, β es el vector de parámetros asociado a $\boldsymbol{x}$ y $c\left(\beta' \boldsymbol{x}\right)$ es una función conocida. Éste modelo es semiparamétrico porque la forma paramétrica es asumida únicamente para el efecto de las covariables mientras que la función de riesgo base es tratada como no paramétrica.
Debido a que $\lambda(t \mid \boldsymbol{x})$ debe ser positivo, comúnmente se emplea a $c\left(\beta' \boldsymbol{x}\right)$ como:

$$c\left(\beta' \boldsymbol{x}\right) = \exp\left(\beta' \boldsymbol{x}\right) = \exp\left(\sum_{i=1}^{n} \beta_i \boldsymbol{x}_i\right) \tag{2.24}$$

y reemplazando éste valor en la ecuación (2.23), se obtiene la expresión dada en la ecuación (1.1).

De lo anterior obtenemos que el logaritmo de $\frac{\lambda(t|\boldsymbol{x})}{\lambda(t)}$ es $\sum_{i=1}^{n} \beta_i \boldsymbol{x}_i$, la cuál es la representación usual de los modelos lineales para efectos de las covariables.

El modelo de Cox es llamado como modelo de riesgos proporcionales porque si se toman dos individuos con conjunto de covariables $\boldsymbol{x}$ y $\boldsymbol{x}^*$, la relación de sus tasas de riesgo es:

$$\frac{\lambda(t \mid \boldsymbol{x})}{\lambda(t \mid \boldsymbol{x}^*)} = \frac{\lambda(t) \exp\left(\sum_{i=1}^{n} \beta_i \boldsymbol{x}_i\right)}{\lambda(t) \exp\left(\sum_{i=1}^{n} \beta_i \boldsymbol{x}_i^*\right)} = \exp\left[\sum_{i=1}^{n} \beta_i \left(\boldsymbol{x}_i - \boldsymbol{x}_i^*\right)\right] \tag{2.25}$$

el cuál es un valor constante. Por lo tanto las tasas de riesgo son proporcionales. El valor obtenido en la ecuación (2.25) es llamado el riesgo relativo (tasa de riesgo) que tiene el evento para un individuo con un factor de riesgo $\boldsymbol{x}$ en comparación al de un individuo con un factor de riesgo $\boldsymbol{x}^*$.

2.1.5.1. Análisis de la función condicional de verosimilitud

Cox al construir su modelo se basa en lo siguiente: "se asume condicionalidad sobre el conjunto de instantes en los cuales ocurren fallas".
Se asume que inicialmente no hay empates (no existen valores que se repiten) y se ordenan los tiempos observados:

$$\left(T_{(1)}, \delta_{(1)}\right), ..., \left(T_{(n)}, \delta_{(n)}\right) \text{ con } T_{(1)} < ... < T_{(n)}$$

Se define: $R\left(T_{(i)}\right)$ como el conjunto de individuos vivos antes del instante $T_{(i)}$ es decir el conjunto riesgo para $T_{(i)}$. En base a la definición de la función de riesgo dada en el modelo de Cox y condicionada al conjunto $R\left(T_{(i)}\right)$ se obtiene (Miller (1981)):

$\mathrm{P}\left[\text{ocurra una muerte en } [T_{(i)}, T_{(i)} + \Delta t[\mid R\left(T_{(i)}\right)\right] = \sum_{j \in R\left(T_{(i)}\right)} e^{\left(\boldsymbol{\beta}' \boldsymbol{x_j}\right)\lambda\left(T_{(i)}\right)}$.

Luego:

$P\left[\text{ocurra la muerte del i-ésimo sujeto en } T_{(i)} \mid \text{ocurrió una muerte en } R\left(T_{(i)}\right)\right] =$

$$= \frac{e^{\boldsymbol{\beta}' \boldsymbol{x_i}} \lambda\left(T_{(i)}\right)}{\sum\limits_{j \in R\left(T_{(i)}\right)} e^{\boldsymbol{\beta}' \boldsymbol{x_j}} \lambda\left(T_{(i)}\right)} = \frac{e^{\boldsymbol{\beta}' \boldsymbol{x_i}}}{\sum\limits_{j \in R\left(T_{(i)}\right)} e^{\boldsymbol{\beta}' \boldsymbol{x_j}}}.$$

La función de verosimilitud que se obtiene es:

$$L_c\left(\boldsymbol{x}, \boldsymbol{\beta}\right) = \prod_{i \in nc} \frac{e^{\boldsymbol{\beta}' \boldsymbol{x_i}}}{\sum\limits_{j \in R\left(T_{(i)}\right)} e^{\boldsymbol{\beta}' \boldsymbol{x_j}}} \tag{2.26}$$

donde nc es el conjunto de observaciones no censuradas. Dado que se desea analizar las fallas o muertes producidas sólo se emplean las observaciones no censuradas.
Cox encuentra sus estimadores con el método de máxima verosimilitud:

$$= \frac{\partial}{\partial \beta} \left\{\log L_c\left(\boldsymbol{x}, \boldsymbol{\beta}\right)\right\} = \left(\frac{\partial}{\partial \beta_1} \log L_c\left(\boldsymbol{x}, \boldsymbol{\beta}\right), ..., \frac{\partial}{\partial \beta_p} \log L_c\left(\boldsymbol{x}, \boldsymbol{\beta}\right)\right) = 0 \tag{2.27}$$

Para hallar la solución de los estimadores para $\boldsymbol{\beta}$ se hace uso del método de Newton, por lo que se requiere hallar:

$$\frac{\partial^2 \log L_c(\boldsymbol{x},\boldsymbol{\beta})}{\partial \beta^2} = \begin{pmatrix} \frac{\partial^2 \log L_c(\boldsymbol{x},\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 \log L_c(\boldsymbol{x},\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L_c(\boldsymbol{x},\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 \log L_c(\boldsymbol{x},\boldsymbol{\beta})}{\partial \beta_p \partial \beta_p} \end{pmatrix} \tag{2.28}$$

Se tiene:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + i^{-1}\left(\widehat{\boldsymbol{\beta}}^{(k)}\right) \frac{\partial}{\partial \boldsymbol{\beta}} \log L_c\left(\boldsymbol{x}, \widehat{\boldsymbol{\beta}}^{(k)}\right), \ \forall\ k \geq 0 \tag{2.29}$$

siendo $\widehat{\boldsymbol{\beta}}^{(0)}$ una aproximación inicial. Si denotamos por:

$$i(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\left\{\log L_c(\boldsymbol{x},\boldsymbol{\beta})\right\}, \tag{2.30}$$

Cox demuestra que el estimador obtenido converge asintóticamente:

$$\widehat{\boldsymbol{\beta}} \xrightarrow{a} N\left(\boldsymbol{\beta}, i^{-1}(\boldsymbol{\beta})\right) \tag{2.31}$$

donde $\xrightarrow{a}$ denota la convergencia en distribución.

2.1.5.2. Estimación de la función de supervivencia asociada al modelo de Cox

A partir de la ecuación (1.1), se hace uso del estimador de $\widehat{\lambda}(t)$ construido por Breslow, para luego estimar a $S(t \mid x)$.

Breslow construye una partición para la variable t: $t_1 < t_2 < ... < t_k$. Las primeras k observaciones distintas corresponden a los k tiempos de muerte distintos (datos no censurados). Se asume que no hay empates. Las $(n-k)$ observaciones restantes (datos censurados), ocuparán los lugares $(k+1)$ a n en cualquier orden.

Se hace uso del modelo de Cox:

$$\log L_c(\boldsymbol{x},\boldsymbol{\beta}) = \sum_{i=1}^{k}\left\{\beta' x_i - \log\left(\sum_{j \in R(T_{(i)})} \exp\left\{\beta' x_j\right\}\right)\right\} \tag{2.32}$$

$$\frac{\partial \log\{L_c(\boldsymbol{x},\boldsymbol{\beta})\}}{\partial \beta_l} = \sum_{i=1}^{k} x_{il} - \frac{\sum_{j \in R(T_{(i)})} x_{jl} \exp\left\{\beta' x_j\right\}}{\sum_{j \in R(T_{(i)})} \exp\left\{\beta' x_j\right\}} = 0,\ 1 \leq l \leq p$$

$$\frac{\partial^2 \log\{L_c(\boldsymbol{x},\boldsymbol{\beta})\}}{\partial \beta_l \beta_m} = -\sum_{i=1}^{k}\left[\frac{\sum_{j \in R(T_{(i)})} x_{jl} x_{jm} \exp\left\{\beta' x_j\right\}}{\sum_{j \in R(T_{(i)})} \exp\left\{\beta' x_j\right\}} + \frac{\sum_{j \in R(T_{(i)})} x_{jl} \exp\left\{\beta' x_j\right\}}{\sum_{j \in R(T_{(i)})} \exp\left\{\beta' x_j\right\}} \frac{\sum_{j \in R(T_{(i)})} x_{jm} \exp\left\{\beta' x_j\right\}}{\sum_{j \in R(T_{(i)})} \exp\left\{\beta' x_j\right\}}\right]$$

Para estimar a $S(t \mid x)$, Breslow estima a $\lambda(t)$ de la siguiente forma:

$$\widehat{\lambda}_i = \frac{1}{(t_i - t_{i-1})} \frac{1}{\sum_{j \in R(T_{(i)})} \exp\left\{\beta' x_j\right\}} \quad \forall\ t \in\ ]t_{i-1}, t_i]. \tag{2.33}$$

La función de supervivencia toma la forma:

$$S(t \mid x) = \exp\{-\Lambda(t \mid \boldsymbol{x})\} \text{ donde: } \Lambda(t \mid \boldsymbol{x}) = \int_0^t \exp\left\{\beta' x\right\} \lambda(u)\, du \tag{2.34}$$

Entonces:

$$S(t \mid x) = \exp\left\{-\exp\left\{\beta' x\right\} \int_0^t \lambda(u)\, du\right\} \tag{2.35}$$

Se estima a $\int_0^t \lambda(u)\, du$ a través de los estimadores de Breslow $(\widehat{\Lambda}_i)$.

$$\widehat{\Lambda}(t) = \int_0^t \widehat{\lambda}(u)\, du$$

Si $t \in [t_l, t_{l+1}[$ entonces tenemos que:

$$\begin{aligned}
\widehat{\Lambda}(t) &= \sum_{i=1}^{l} \int_{t_{i-1}}^{t_i} \widehat{\lambda}_i du + \int_{t_l}^{t} \widehat{\lambda}_l du \\
&= \sum_{i=1}^{l} \widehat{\lambda}_i (t_i - t_{i-1}) + \widehat{\lambda}_l (t - t_l) \\
&= \sum_{i=1}^{l} \frac{1}{(t_i - t_{i-1})} \frac{1}{\sum_{j \in R(T_{(i)})} \exp\left\{\widehat{\beta}' x_j\right\}} (t_i - t_{i-1}) + \frac{1}{(t_l - t_{l-1})} \frac{1}{\sum_{j \in R(T_{(i)})} \exp\left\{\widehat{\beta}' x_j\right\}} (t - t_l) \\
&= \sum_{i=1}^{l} \frac{1}{\sum_{j \in R(T_{(i)})} \exp\left\{\widehat{\beta}' x_j\right\}} + \frac{(t - t_l)}{(t_l - t_{l-1})} \frac{1}{\sum_{j \in R(T_{(i)})} \exp\left\{\widehat{\beta}' x_j\right\}}.
\end{aligned}$$

La estimación de la función de supervivencia es:

$$\begin{aligned}
\widehat{S}_L(t \mid x) &= \exp\left\{-\exp\left\{\widehat{\beta}' x\right\} \widehat{\Lambda}(t)\right\} \\
&= \exp\left\{-\exp\left\{\widehat{\beta}' x\right\} \left(\sum_{i=1}^{l} \left(\frac{1}{\sum_{j \in R(T_{(i)})} \exp\left\{\widehat{\beta}' x_j\right\}} \right) + \frac{(t - t_l)}{(t_l - t_{l-1})} \frac{1}{\sum_{j \in R(T_{(i)})} \exp\left\{\widehat{\beta}' x_j\right\}} \right) \right\}
\end{aligned}$$

definida $\forall t \leq t_k$ ya que si $t > t_k$, $\lambda(t \mid x)$ no está definida.

El estimador de Breslow para $S(t \mid x)$ toma la forma:

$$\widehat{S}_B(t \mid x) = \left\{ \prod_{i=1}^{l} \left(1 - \frac{1}{\sum_{j \in R(T_{(i)})} \exp\left\{\widehat{\beta}' x_j\right\}} \right) \right\}^{\exp\left\{\beta' x\right\}} \tag{2.36}$$

2.2. Intervalos de Confianza

En esta parte se desarrollan los conceptos de los intervalos de confianza para luego introducir a los del bootstrap.

2.2.1. Test e intervalos de confianza

Para el censuramiento del Tipo I y para el censuramiento aleatorio, bajo condiciones de suavización se tiene a la ecuación (2.31).
Usualmente para el Tipo II de censuramiento, éste resultado también se mantiene pero las proposiciones son diferentes.
Para la prueba $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^0$ o para la construcción de los intervalos de confianza se tienen tres procedimientos,

1. Método de Wald

$$\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)' i(\boldsymbol{\beta}^0) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right) \overset{a}{\sim} \chi^2_p \text{ bajo } H_0, \tag{2.37}$$

   donde la matriz $i(\boldsymbol{\beta})$ se definió en la ecuación (2.30).
   Como alternativa se puede reemplazar $i\left(\widehat{\boldsymbol{\beta}}\right)$ por $i\left(\boldsymbol{\beta}^0\right)$ donde $\boldsymbol{\beta}^0$ es una estimación inicial del vector de parámetros.

2. Método de Neyman-Pearson y Wilks

$$-2 \log \frac{L\left(\boldsymbol{\beta}^0\right)}{L\left(\widehat{\boldsymbol{\beta}}\right)} \overset{a}{\sim} \chi^2_p \text{ bajo } H_0. \tag{2.38}$$

3. Método de Rao

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log L\left(\boldsymbol{\beta}^0\right)' i^{-1}\left(\boldsymbol{\beta}^0\right) \frac{\partial}{\partial \boldsymbol{\beta}} \log L\left(\boldsymbol{\beta}^0\right) \overset{a}{\sim} \chi^2_p \text{ bajo } H_0. \tag{2.39}$$

Notar que el método de Rao no usa el estimador de máxima verosimilitud, por consiguiente el cálculo iterativo no es necesario. Sin embargo, usualmente se desean estimar intervalos de confianza; por lo tanto se necesita calcular $\widehat{\boldsymbol{\beta}}$ de todos modos. Una vez que se tenga $\widehat{\boldsymbol{\beta}}$ y $i\left(\boldsymbol{\beta}^0\right)$, el método de Wald es fácil de calcular.

2.2.2. El método Delta

Suponga que la variable aleatoria Y tiene media μ y varianza σ^2 y que se desea la distribución de la función $g(Y)$ para lo cuál se construye un desarrollo de Taylor para $g(Y)$ en torno de μ, asumiendo de que se satisfacen las hipótesis necesarias, se obtiene (Miller (1981)):

$$g(Y) = g(\mu) + (Y - \mu) g'(\mu) + (Y - \mu)^2 g''(\mu) + ... \tag{2.40}$$

Calculando la $E[g(Y)]$ y $V[g(Y)]$ y prescindiendo de los términos de orden superior:

$$\begin{aligned} E\left[g\left(Y\right)\right] &= E\left[g\left(\mu\right)+\left(Y-\mu\right)g^{'}\left(\mu\right)\right] \\ &= E\left[g\left(\mu\right)\right]+E\left[\left(Y-\mu\right)g^{'}\left(\mu\right)\right] \\ &= g\left(\mu\right) \end{aligned}$$

$$\begin{aligned} V\left[g\left(Y\right)\right] &= V\left[g\left(\mu\right)+\left(Y-\mu\right)g^{'}\left(\mu\right)\right] \\ &= V\left[g\left(\mu\right)\right]+V\left[\left(Y-\mu\right)g^{'}\left(\mu\right)\right] \\ &= \sigma^{2}\left(g^{'}\left(\mu\right)\right)^{2} \end{aligned}$$

con lo que se consigue:

$$g\left(Y\right)\approx\left(g\left(\mu\right),\sigma^{2}\left(g^{'}\left(\mu\right)\right)^{2}\right) \tag{2.41}$$

donde $\approx$ significa que está aproximadamente distribuido.

Si además $Y\overset{a}{\sim}N\left(\mu,\sigma^{2}\right)$ y g es continua, se sabe que:

$$g\left(Y\right)\overset{a}{\sim}N\left(g\left(\mu\right),\sigma^{2}\left(g^{'}\left(\mu\right)\right)^{2}\right). \tag{2.42}$$

donde $\overset{a}{\sim}$ significa que está asintóticamente distribuido.

El método delta también tiene una versión multivariada. Suponga que $\left(\begin{smallmatrix}x\\y\end{smallmatrix}\right)$ es un vector aleatorio con $E\left[\left(\begin{smallmatrix}x\\y\end{smallmatrix}\right)\right]=0$ y $V\left[\left(\begin{smallmatrix}x\\y\end{smallmatrix}\right)\right]=0$, entonces:

$$\begin{pmatrix}X\\Y\end{pmatrix}\sim\left(\begin{pmatrix}\mu_{x}\\\mu_{y}\end{pmatrix},\begin{pmatrix}\sigma_{x}^{2}&\sigma_{xy}\\\sigma_{xy}&\sigma_{y}^{2}\end{pmatrix}\right), \tag{2.43}$$

y que se desea la distribución $g\left(X,Y\right)$. Entonces:

$$g\left(X,Y\right)=g\left(\mu_{x},\mu_{y}\right)+\left(X-\mu_{x}\right)\frac{\partial}{\partial x}g\left(\mu_{x},\mu_{y}\right)+\left(Y-\mu_{y}\right)\frac{\partial}{\partial y}g\left(\mu_{x},\mu_{y}\right)+..., \tag{2.44}$$

por lo tanto:

$$g\left(X,Y\right)\approx\left(g\left(\mu_{x},\mu_{y}\right),\sigma_{x}^{2}\left(\frac{\partial}{\partial x}g\right)^{2}+2\sigma_{xy}\frac{\partial}{\partial x}g\frac{\partial}{\partial y}g+\sigma_{y}^{2}\left(\frac{\partial}{\partial y}g\right)^{2}\right). \tag{2.45}$$

Si además $\left(X,Y\right)\overset{a}{\sim}$ Normal, entonces $g\left(X,Y\right)\overset{a}{\sim}$ Normal en los parámetros de (2.45).
El método delta es muy útil, por ejemplo para obtener un valor aproximado de Var $\left(\overline{X}/\overline{Y}\right)$ o Var $\left(\overline{X}\ \overline{Y}\right)$.

2.2.3. Intervalos de confianza para la función de supervivencia

El estimador de Producto Límite proporciona una estimación que resume la experiencia de mortalidad de una población dada. El correspondiente error estándar proporciona información limitada acerca de la precisión de la estimación. Se utilizan estos estimadores para proveer intervalos de confianza para la función de supervivencia en un tiempo fijo t_0. Los intervalos son construidos para asegurar, con un nivel de confianza dado $1-\alpha$ que el verdadero valor de la función de supervivencia en un momento t_0, cae en el intervalo que se va a construir.

Sea $\sigma_S^2(t) = \widehat{V}\left[\widehat{S}(t)\right]/\widehat{S}^2(t)$ donde $\sigma_S^2(t)$ es la fórmula de Greenwood.

El intervalo de confianza comúnmente usado al nivel $100\times(1-\alpha)$ % para la función de supervivencia en el tiempo t_0, denominado intervalo de confianza lineal, esta definido por:

$$\left[\widehat{S}(t_0) - Z_{1-\alpha/2}\sigma_S(t_0)\,\widehat{S}(t_0), \widehat{S}(t_0) + Z_{1-\alpha/2}\sigma_S(t_0)\,\widehat{S}(t_0)\right] \tag{2.46}$$

donde $Z_{1-\alpha/2}$ es el $(1-\alpha/2)$ percentil de la distribución normal estándar.

Mejores intervalos de confianza pueden ser construidos transformando primero $\widehat{S}(t_0)$. Estos estimadores mejorados fueron propuestos por Borgan y Liestol (1990). La primera transformación sugerida es la transformación logarítmica de la tasa de riesgo acumulada. El intervalo de confianza al nivel $100\times(1-\alpha)$ % log-transformado para la función de supervivencia en t_0 esta dado por:

$$\left[\widehat{S}(t_0)^{1/\theta}, \widehat{S}(t_0)^{\theta}\right], \text{donde } \theta = \exp\left\{\frac{Z_{1-\alpha/2}\sigma_S(t_0)}{\log\left[\widehat{S}(t_0)\right]}\right\} \tag{2.47}$$

Notar que este intervalo no es simétrico cerca de la estimación de la función de supervivencia.

La segunda transformación es la arcoseno-raíz cuadrada de la función de supervivencia que brinda el siguiente intervalo de confianza al nivel $100\times(1-\alpha)$ % para la función de supervivencia:

$$a \leq S(t_0) \leq b \tag{2.48}$$

donde:

$$a = \operatorname{sin}^2\left\{\text{máx}\left[0, \arcsin\left(\widehat{S}(t_0)^{1/2}\right) - 0.5Z_{1-\alpha/2}\sigma_S(t_0)\left(\frac{\widehat{S}(t_0)}{1-\widehat{S}(t_0)}\right)^{1/2}\right]\right\}$$

y

$$b = \operatorname{sin}^2\left\{\text{mín}\left[\frac{\pi}{2}, \arcsin\left(\widehat{S}(t_0)^{1/2}\right) + 0.5Z_{1-\alpha/2}\sigma_S(t_0)\left(\frac{\widehat{S}(t_0)}{1-\widehat{S}(t_0)}\right)^{1/2}\right]\right\}$$

Por ejemplo para encontrar el intervalo de confianza log-transformado del 95 % para la función de supervivencia de un año con $\sigma_S(t_0) = 0.1479$ y $\widehat{S}(t_0) = 0.5492$, se encuentra que

$\theta = \exp\left[\frac{1.96\times 0.1479}{\log(0.5492)}\right] = 0.6165$, por lo tanto el intervalo es $\left(0.54921^{1/0.6165}, 0.5492^{0.6165}\right) = (0.3783, 0.6911)$.

Y para el intervalo de confianza arcoseno-raíz cuadrada transformada con los mismos valores expuestos en el párrafo anterior es $a \leq S(t_0) \leq b$ con:

$$a = \sin^2\left\{\text{máx}\left[0, \arcsin\left(0.5492\right)^{1/2}\right) - 0.5 \times 1.96 \times 0.1479 \times \left(\frac{0.5492}{1-0.5492}\right)^{1/2}\right]\right\}$$

$$b = \sin^2\left\{\text{mín}\left[\frac{\pi}{2}, \arcsin\left(0.5492\right)^{1/2}\right) + 0.5 \times 1.96 \times 0.1479 \times \left(\frac{0.5492}{1-0.5492}\right)^{1/2}\right]\right\}$$

$$= (0.3903, 0.7032)$$

Notar que:

1. Bie et al. (1987) presentó los intervalos de confianza $100 \times (1-\alpha)$ % para la función de riesgo acumulada. Similares intervalos de confianza fueron construidos para la función de supervivencia, aquí se muestran tres posibles intervalos que corresponden a tres transformaciones de la función de riesgo acumulada. Los intervalos son:

   Lineal: $\left[\widetilde{\Lambda}(t_0) - Z_{1-\alpha/2}\sigma_\Lambda(t_0), \widetilde{\Lambda}(t_0) + Z_{1-\alpha/2}\sigma_\Lambda(t_0)\right]$ (2.49)

   Log-transformado: $\left[\widetilde{\Lambda}(t_0)/\phi, \phi\widetilde{\Lambda}(t_0)\right]$ donde $\phi = \exp\left[\frac{Z_{1-\alpha/2}\sigma_\Lambda(t_0)}{\widetilde{\Lambda}(t_0)}\right]$ (2.50)

   Arcoseno-raíz cuadrada transformado: $a \leq \Lambda(t_0) \leq b$ con: (2.51)

   $$a = -2\log\left\{\sin\left[\text{mín}\left(\frac{\pi}{2}, \arcsin\left[\exp\left\{-\widetilde{\Lambda}(t_0)/2\right\}\right] + 0.5Z_{1-\alpha/2}\sigma_\Lambda(t_0)\left\{\exp\left\{\widetilde{\Lambda}(t_0)\right\} - 1\right\}^{-1/2}\right)\right]\right\}$$

   $$a = -2\log\left\{\sin\left[\text{mín}\left(\frac{\pi}{2}, \arcsin\left[\exp\left\{-\widetilde{\Lambda}(t_0)/2\right\}\right] + 0.5Z_{1-\alpha/2}\sigma_\Lambda(t_0)\left\{\exp\left\{\widetilde{\Lambda}(t_0)\right\} - 1\right\}^{-1/2}\right)\right]\right\}$$

2. Borgan y Liestol (1990) mostraron que tanto los intervalos de confianza log-transformados como los arcoseno-raíz cuadrada transformados para S tienen un mejor rendimiento que el usual intervalo de confianza lineal. Ambos dan una probabilidad de cobertura para un intervalo del 95 % para muestras tan pequeñas como de 25 datos que tienen como mucho 50 % de censura, excepto en el extremo del lado derecho en donde habrán poco datos. El tamaño de muestra necesario para que el intervalo de confianza lineal estándar tenga la probabilidad de cobertura correcta es mucho más grande.

   Para muestras muy pequeñas, el intervalo arcoseno-raíz cuadrada transformado tiende a ser un poco conservador en la medida en que la probabilidad de cobertura es un poco mayor que $(1-\alpha)$, mientras que para el intervalo log-transformado, la probabilidad de cobertura es un poco más pequeña que $(1-\alpha)$. La probabilidad de cobertura para el intervalo lineal en estos casos es mucho más pequeña que $(1-\alpha)$. Observaciones

similares son hechas por Bie et al. (1987) para la estimación del intervalo de la tasa de riesgo acumulada. Para muestras muy grandes, los tres métodos son equivalentes.

3. Intervalos de confianza alternativos para la tasa de riesgo acumulada pueden ser encontrados tomando el logaritmo natural de los intervalos de confianza construidos para la función de supervivencia. Similarmente la exponencial de los límites de confianza para el riesgo acumulado permiten obtener un intervalo de confianza para la función de supervivencia.

4. Tanto los intervalos de confianza log-transformado y los arcoseno-raíz cuadrada transformado son diferentes al intervalo lineal, y no son simétricos cerca del estimador puntual de la función de supervivencia o de la tasa de riesgo acumulada. Esto es apropiado para muestras pequeñas donde los estimadores puntuales son parciales y la distribución de los estimadores es sesgada.

5. Los intervalos de confianza expuestos son válidos únicamente en el punto t_0. Un uso común incorrecto de éstos es graficar a todos los valores de t e interpretar las curvas obtenidas como una banda de confianza; esto es, estas curvas son interpretadas como que tienen por ejemplo, 95 % de confianza de que la función de supervivencia se encuentre íntegramente dentro de la banda.

6. La construcción de intervalos de confianza lineales se deriva de la normalidad asintótica del Producto Límite o de los estimadores de Nelson-Aalen.

7. El intervalo log-transformado fue propuesto por Kalbfleisch y Prentice (1980) y el arcoseno-raíz cuadrada transformado por Nair (1984)).

8. El intervalo de confianza log-transformado se basa en encontrar un intervalo de confianza para el logaritmo de la función de riesgo acumulada. Éste es algunas veces llamado el intervalo log-log transformado desde que la función de riesgo acumulada es el logaritmo negativo de la función de supervivencia.

2.3. Método Bootstrap

Bootstrap es un método para asignar medidas de precisión a las estimaciones de la muestra. Esta técnica permite la estimación de la distribución muestral de casi cualquier estadística utilizando métodos de re-muestreo muy simples (Efron y Tibshirani (1993)).
El principio clave del bootstrap es proporcionar una manera de simular repetidas observaciones de una población desconocida utilizando la muestra obtenida como base.
También puede utilizarse para construir pruebas de hipótesis. A menudo se utiliza como una alternativa a la inferencia basada en supuestos paramétricos cuando éstos están en duda, o donde la inferencia paramétrica o es imposible o requiere fórmulas muy complicadas para el cálculo del error estándar.
Para una aproximación de la distribución, una de las opciones es el uso de la distribución empírica de los datos observados. En el caso de un conjunto de observaciones provenientes

de una población i.i.d., puede implementarse mediante la construcción de un número de muestreos del conjunto de datos observado, cada uno de los cuales se obtiene por muestreo aleatorio con reemplazo del conjunto de datos original.
Una gran ventaja del bootstrap es su simplicidad. Es sencillo obtener estimaciones del error estándar y de los intervalos de confianza para complejos estimadores de los parámetros de la distribución, como percentiles, proporciones, odds-ratio y coeficientes de correlación. Además, es una forma adecuada para controlar y verificar la estabilidad de los resultados.
Por su parte una desventaja es que aunque bajo ciertas condiciones, es asintóticamente consistente, no proporciona garantías para muestras finitas. Además, tiene tendencia a ser demasiado optimista. La aparente sencillez puede ocultar el hecho de que se hacen importantes suposiciones al realizar el análisis bootstrap (por ejemplo, la independencia de las muestras).
Las distribuciones de los estadísticos de prueba, estimadores o cantidades pivotales a menudo se pueden calcular mediante simulación. Entiéndase por cantidad pivotal como una función de las observaciones y de los parámetros no observables cuya distribución de probabilidad no depende de parámetros desconocidos.
La metodología Bootstrap hace uso de medios similares a los de la simulación pero en un contexto más general donde la aproximación relativamente cercana de una distribución no es necesariamente posible.
Los métodos para estimar intervalos de confianza son los siguientes (Lawless (2003)):

1. Bootstrap Paramétrico: Considerar una muestra independiente e idénticamente distribuida de un modelo paramétrico $Y \sim F(y;\theta)$. Intervalos de confianza para un parámetro especifico ξ están basados en cantidades pivotales. Sin embargo, para la mayoría de los modelos de cantidades pivotales e intervalos de confianza con probabilidades de cobertura prescrita exacta, α no existe. En ese caso, se construyen intervalos de confianza que se basan en cantidades $W = g(y_1, y_2, ..., y_n; \xi)$ que son asintóticamente pivotales: la distribución límite de W cuando $n \to \infty$ no depende de θ. En este caso, los intervalos de confianza obtenidos por probabilidades invertidas para W, basadas en su distribución límite, típicamente tienen probabilidad de cobertura $\alpha + C(\theta) n^{-1/2}$, con $C(\theta) = ni(\theta)^{-1}$ de acuerdo con la teoría de máxima verosimilitud y donde $i(\theta)$ es la matriz de información de Fisher. Los pivotales aproximados frecuentemente usados son los de la teoría asintótica máxima verosímil: la aproximación $N(0,1)$,

$$W = \left(\widehat{\xi} - \xi\right) / se\left(\widehat{\xi}\right). \tag{2.52}$$

   El método bootstrap paramétrico básico para obtener intervalos de confianza de una cantidad pivotal asintótica $W = g(y_1, y_2, ..., y_n; \xi)$ es como sigue.
   Sea $\widehat{\theta}$ el estimador de máxima verosimilitud de θ en el modelo asumido $F(y;\theta)$ basado en una muestra $y_1, y_2, ..., y_n$. A continuación se llevan a cabo los siguientes pasos:

   a) Generar una muestra bootstrap pseudoaleatoria $y_1^*, y_2^*, ..., y_n^*$ de $F\left(y;\widehat{\theta}\right)$, con lo que se obtiene el valor:

$$w^* = g\left(y_1^*, y_2^*, ..., y_n^*; \widehat{\xi}\right), \tag{2.53}$$

donde $\widehat{\xi}$ es el estimador de máxima verosimilitud de ξ basado en $\widehat{\theta}$.

b) Repetir este proceso B veces, produciendo valores $w_1^*, w_2^*, ..., w_B^*$. La distribución de W puede ser estimada de $w_1^*, w_2^*, ..., w_B^*$.
El q-ésimo cuantil para W es estimado por $w_{(qB)}^*$, donde asumimos por simplicidad que qB es un número entero, tal que $w_{(qB)}^*$ es el (qB)-ésimo valor más pequeño entre los w_j^*. Entonces para $q_2 > q_1$ se tiene,

$$P\left(w_{(q_1 B)}^* \leq W \leq w_{(q_2 B)}^*\right) = q_2 - q_1 \tag{2.54}$$

y esto puede ser invertido para dar un intervalo de confianza aproximado al $(q_2 - q_1)$ % para ξ.

Por ejemplo, se puede suponer que se conoce el valor de W en la ecuación (2.52). Entonces cada muestra bootstrap $y_1^*, y_2^*, ..., y_n^*$ da una estimación $\widehat{\xi}^*$ y un error estándar $se\left(\widehat{\xi}^*\right)$ para ξ, y el valor

$$w^* = \left(\widehat{\xi}^* - \widehat{\xi}\right) / se\left(\widehat{\xi}^*\right). \tag{2.55}$$

La probabilidad dada en (2.54) nos brinda el siguiente intervalo de confianza

$$\widehat{\xi} - {w^*}_{(q_2 B)} se\left(\widehat{\xi}\right) \leq \xi \leq \widehat{\xi} + {w^*}_{(q_1 B)} se\left(\widehat{\xi}\right). \tag{2.56}$$

Éste método es frecuentemente llamado el bootstrap studentizado o el método bootstrap-t. Las cantidades W de (2.52) y $\Lambda(\xi)$ tienen distribuciones límite $N(0,1)$ y ${\chi^2}_{(1)}$ respectivamente, y para un n grande, las estimaciones bootstrap basadas en ${w^*}_1, {w^*}_2, ..., {w^*}_B$ reflejarán esto. La aproximación bootstrap es usualmente mejor para pequeños valores de n, aunque un valor grande de B debe ser necesario cuando q_1 es cercano a cero o q_2 es cercano a uno. También es importante para la exactitud de los intervalos de confianza que la cantidad W en que se basa debe ser lo más cercana posible a la pivotal. Con cantidades de la forma (2.52) es aconsejable usar una parametrización que facilite esto. Parametrizaciones o transformaciones de ξ que estabilizan la varianza de $\widehat{\xi}$ o reducen la asimetría de $\Lambda(\xi)$ pueden mejorar sustancialmente la precisión de los intervalos de confianza. Procedimientos basados en (2.52), a diferencia de los basados en $\Lambda(\xi)$ no son invariantes.

Alrededor de $B = 2,000$ muestras bootstrap son frecuentemente sugeridas para una buena precisión, y $5,000$ o más cuando $q_2 - q_1$ es cercano a 1 (Lawless (2003)). Sin embargo, valores grandes de B no pueden superar la inexactitud debido a que la distribución de W depende de θ.

El bootstrap paramétrico también puede ser usado para revisar la precisión de aproximaciones de distribuciones límites como la normalidad o la χ^2 asintótica para los cuantiles pivotales a través de gráficos de probabilidad de w_j^*, o comparando las probabilidades o cuantiles estimados con su contraparte normal o χ^2.

2. Bootstrap No Paramétrico: En algunas aplicaciones no es posible generar valores y muestras bootstrap del tipo dado en (2.53) de un modelo paramétrico. Algunas veces es difícil simular datos de éste modelo, pero es más frecuente que el proceso de generación de los datos no esté totalmente especificado. Por ejemplo, muestras con censura $y_i = (t_i, \delta_i)\,,\ i = 1, 2, ..., n$ a menudo surgen bajo un proceso con censura aleatoria que no es conocido.
El método bootstrap no paramétrico reemplaza el primer paso en el algoritmo dado en (2.53) por el siguiente:

   a) Generar una muestra bootstrap $y_1^*, y_2^*, ..., y_n^*$ a través de extraer aleatoriamente n ítems con reemplazo, de $y_1, y_2, ..., y_n$.

   El resto del procedimiento bootstrap se mantiene como el anterior. Bajo ciertas condiciones, éste proceso provee aproximaciones asintóticas precisas a distribuciones de variables como con W en (2.52) o con la correspondiente estadística verosímil $\Lambda(\xi)$.

   Existe otro procedimiento bootstrap no paramétrico pero el anterior es más fácil de usar y generaliza a otros problemas relacionados con elementos de datos independientes $d_1, d_2, ..., d_n$. Así por ejemplo, si y_i tiene un vector de covariables asociado $\boldsymbol{x_i}$ y está sujeto a censura por la derecha, entonces las muestras bootstrap pueden ser obtenidas seleccionando muestras de tamaño n con reemplazo de $d_i = (y_i, \delta_i, \boldsymbol{x_i})\,,\ i = 1, 2, ..., n$.

   Por ejemplo, se considera una estimación no paramétrica de una función de supervivencia $S(t)$ sobre la base de una muestra aleatoria con censura $(t_i, \delta_i)\,,\ i = 1, 2, ..., n$. La cantidad pivotal aproximada puede ser usada para obtener intervalos de confianza para $S(t)$. Luego se generan muestras bootstrap individuales seleccionando n ítems $(t_i^*, \delta_i^*)\,,\ i = 1, 2, ..., n$ con reemplazo de $\{(t_i, \delta_i), i = 1, 2, ..., n\}$. Cada muestra da un estimador Kaplan-Meier $\widehat{S}^*(t)$ asociado al error estándar $\widehat{\sigma}_S^*(t)$, con lo que se obtiene:

$$Z_1^* = \frac{\widehat{S}^*(t) - \widehat{S}(t)}{\widehat{\sigma}_S^*(t)} \tag{2.57}$$

   donde $\widehat{S}(t)$ es el estimador de Kaplan-Meier de los datos $(t_i, \delta_i)\,,\ i = 1, 2, ..., n$. La generación de B muestras boostrap y de sus valores correspondientes Z_1^* proveen una estimación empírica de los cuantiles q_1 y q_2. Esto da un intervalo de confianza al $(q_2 - q_1)$ % para $S(t)$, exactamente como en (2.54) y (2.56).

Respecto a la metodología bootstrap, en Lawless (2003) se menciona que:

1. Bootstrap trabaja para funciones suavizadas de los datos, y es muy útil para problemas donde la teoría asintótica de máxima verosimilitud no es fácil de aplicar.

2. Se han desarrollado una variedad de métodos diseñados para mejorar la precisión de aproximaciones bootstrap.

3. En algunos problemas es convencional hacer inferencias condicionales en ciertos aspectos de los datos. En particular, para covariables se acostumbra a elegir los valores observados x_i $(i = 1, 2, ..., n)$. Métodos bootstrap paramétricos hacen esto, pero el método no paramétrico descrito previamente no, el cuál considera observaciones (y_i, x_i) como aleatorias. Métodos no paramétricos que condicionan a los x_i también han sido propuestos. En la práctica hay una pequeña diferencia entre los intervalos de confianza bootstrap desarrollados bajo un marco x fijo y uno aleatorio.

2.4. Teoría asintótica

A continuación se presentan los conceptos ligados con esta teoría:

2.4.1. Convergencia asintótica

2.4.1.1. Tipos de convergencia asintótica

Se presentan tres tipos de convergencia asintótica que usualmente se emplean en la literatura: la convergencia fuerte o casi segura, la convergencia en probabilidad y la convergencia en distribución o débil.

Sea $(Y_n)_{n\in N^+}$ una secuencia de variables aleatorias definidas en un mismo espacio de probabilidad $(\Omega, \digamma, P)$, donde Ω es el conjunto de posibles resultados de un experimento aleatorio, $\digamma$ es una σ-álgebra de Ω y P es la probabilidad para los elementos de $\digamma$. Además Y es una variable aleatoria definida en el mismo espacio.

Se dice que $(Y_n)_{n\in N^+}$ converge a Y fuertemente o casi seguramente, si se cumple:

$$P\left(\lim_{n\to\infty} Y_n = Y\right) = 1. \tag{2.58}$$

Esta última expresión se denota por $Y_n \overset{c.s.}{\to} Y$.

En el caso de la convergencia en probabilidad, se debe cumplir la siguiente expresión $\forall\ \epsilon > 0$:

$$\lim_{n\to\infty} P\left(|Y_n - Y| > \epsilon\right) = 0. \tag{2.59}$$

Esto se denota por $Y_n \overset{P}{\to} Y$.

Para la convergencia en distribución o débil, se debe considerar que la secuencia de variables aleatorias $(Y_n)_{n\in N^+}$ tiene funciones de distribución acumuladas $F_1, F_2, ...$ respectivamente. Se dice que la secuencia $(Y_n)_{n\in N^+}$ converge en distribución a la variable aleatoria Y, con distribución acumulada F, si para todo y, punto de continuidad de F, se tiene que:

$$\lim_{n\to\infty} F_n\left(y\right) = F\left(y\right). \tag{2.60}$$

Este último tipo de convergencia se denota como $Y_n \overset{D}{\to} Y$.

Cabe destacar que la convergencia casi segura implica la convergencia en probabilidad; es decir, si $Y_n \xrightarrow{c.s.} Y \Rightarrow Y_n \xrightarrow{P} Y$. Asimismo, la convergencia en probabilidad involucra la convergencia en distribución: $Y_n \xrightarrow{P} Y \Rightarrow Y_n \xrightarrow{D} Y$.

2.4.1.2. Consistencia asintótica

La secuencia de estimadores $(\widehat{\boldsymbol{\gamma}}_n)_{n\in N+}$ es débilmente consistente para el parámetro $\boldsymbol{\gamma}$, si $\widehat{\boldsymbol{\gamma}_n} \xrightarrow{P} \boldsymbol{\gamma}$.
En el caso de que se trate de la consistencia casi segura $\widehat{\boldsymbol{\gamma}_n} \xrightarrow{c.s.} \boldsymbol{\gamma}$, se dice que el estimador es fuertemente consistente.
A continuación se presentan algunas propiedades de la consistencia de los estimadores:

1. Si $\widehat{\boldsymbol{\gamma}_n} \xrightarrow{c.s.} \boldsymbol{\gamma} \Rightarrow \widehat{\boldsymbol{\gamma}_n} \xrightarrow{P} \boldsymbol{\gamma}$.
2. Si $\lim_{n\to\infty} E\left(\widehat{\boldsymbol{\gamma}_n}\right) = \boldsymbol{\gamma}$ y $\lim_{n\to\infty} V\left(\widehat{\boldsymbol{\gamma}_n}\right) = 0 \Rightarrow \widehat{\boldsymbol{\gamma}_n} \xrightarrow{P} \boldsymbol{\gamma}$, donde $V(\cdot)$ es la varianza de $\widehat{\boldsymbol{\gamma}_n}$.
3. Si $\widehat{\boldsymbol{\gamma}_n} \xrightarrow{c.s.} \boldsymbol{\gamma} \Rightarrow g\left(\widehat{\boldsymbol{\gamma}_n}\right) \xrightarrow{c.s.} g(\boldsymbol{\gamma})$, donde $g(\cdot)$ es una función continua.

2.4.1.3. Teoría del Límite Central

Según el Teorema del Límite Central, si se tiene para cualquier secuencia de variables aleatorias independientes e idénticamente distribuidas $(X_n)_{n\in N^+}$, con $E(X_n) = \mu$ y $V(X_n) = \sigma^2$, se tiene;

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \xrightarrow{D} N(0,1). \tag{2.61}$$

Del teorema anterior, se desprende que $\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{D} N\left(0, \sigma^2\right)$.
Asimismo, también se puede demostrar que si $(X_n)_{n\in N^+}$ es una secuencia de variables aleatorias i.i.d., con media μ y varianza σ^2, entonces:

$$Z_n = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \xrightarrow{D} N(0,1) \tag{2.62}$$

donde $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.
De esta manera, la variable aleatoria Z se distribuye asintóticamente como una normal estándar.

2.4.1.4. Distribución asintótica del estimador de máxima verosimilitud

Se dice que la familia de distribuciones o modelos de densidad o de probabilidad $\{f(\boldsymbol{x} \mid \boldsymbol{\beta})\}_{\boldsymbol{\beta}\in B}$, donde B es el espacio paramétrico del modelo, satisface las condiciones de regularidad, si se verifican las siguientes suposiciones:

1. B es un conjunto abierto.

2. Las distribuciones $f(\boldsymbol{x} \mid \boldsymbol{\beta})$ tienen el mismo soporte para todo $\boldsymbol{\beta} \in B$, es decir, $A = \{x : f(\boldsymbol{x} \mid \boldsymbol{\beta}) \geq 0\}$ es independiente de $\boldsymbol{\beta}$.

3. $\forall x \in A$, $f(\boldsymbol{x} \mid \boldsymbol{\beta})$ es tres veces derivable con respecto a $\boldsymbol{\beta}$, la tercera derivada es continua en $\boldsymbol{\beta}$ y $\int f(\boldsymbol{x} \mid \boldsymbol{\beta})\, dx$ puede ser derivado tres veces bajo el signo de la integral.

4. $\forall\ \boldsymbol{\beta_0} \in B$, existe un número positivo c y una función $M(\boldsymbol{x})$ (que pueden depender de $\boldsymbol{\beta_0}$), tal que:

$$\left|\frac{\partial^3}{\partial \boldsymbol{\beta}^3} \log f(\boldsymbol{x}|\boldsymbol{\beta})\right| \leq M(\boldsymbol{x}),\ \forall x \in A, \boldsymbol{\beta_0} - c < \boldsymbol{\beta} < \boldsymbol{\beta_0} + c,$$

con $E_{\boldsymbol{\beta_0}}[M(X)] < \infty$, donde $E_{\boldsymbol{\beta_0}}[M(X)] = \int_{M(X)} f(x/\beta_0)\, x dx$.

Con ello, se obtiene un resultado asintótico para la distribución de los estimadores obtenidos a través de los métodos de máxima verosimilitud. Si el modelo probabilístico $\{f(\boldsymbol{x} \mid \boldsymbol{\beta})\}_{\boldsymbol{\beta} \in B}$ satisface las condiciones de regularidad, X es una variable aleatoria con función de distribución $f(\boldsymbol{x} \mid \boldsymbol{\beta})$, con $\boldsymbol{\beta} \in B$ y el tamaño de muestra n es grande, entonces el estimador de máxima verosimilitud $\widehat{\boldsymbol{\beta}}$ para $\boldsymbol{\beta}$ tiene una distribución asintótica normal de media $\boldsymbol{\beta}$ y varianza:

$$V\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{n} E\left[-\frac{\partial^2 \log f(\boldsymbol{x} \mid \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right]^{-1}. \tag{2.63}$$

Como se puede observar, la varianza asintótica también se puede reexpresar utilizando la información de Fisher $i(\boldsymbol{\beta}) = E\left[-\frac{\partial^2 \log f(\boldsymbol{x}|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right]$, de la siguiente manera:

$$V\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{n} i(\boldsymbol{\beta})^{-1} \tag{2.64}$$

Capítulo 3

Intervalos de confianza para la mediana de supervivencia

Basados en la investigación de Tze y Zheng (2006), en esta sección se detalla la prueba basada en intervalos de confianza para el p-ésimo cuantil $\xi_p(\boldsymbol{x})$ dado el vector de covariables $\boldsymbol{x}$ en el modelo de Cox, y también se proporciona un algoritmo asociado para calcular los puntos extremos del intervalo.

Como se ha mencionado en la sección (1.1), el modelo de riesgos proporcionales de Cox (1972) es un modelo de regresión log-lineal.
En muchas aplicaciones, es útil estimar la mediana de supervivencia dado el vector de covariables de los sujetos. En particular, mediante la combinación de $\widehat{\boldsymbol{\beta}}$ con la estimación de la función de riesgo acumulada subyacente de Breslow (1974) $(\widehat{\Lambda})$, Miller y Halpern (1982) usaron la mediana de la función de distribución $1-\exp\left\{-\widehat{\Lambda}(.)\, e^{\widehat{\boldsymbol{\beta}}'\boldsymbol{x}}\right\}$ para estimar la mediana de supervivencia. Dabrowska y Doksum (1987) y posteriormente Burr y Doss (1993) estudiaron el problema de la construcción de intervalos de confianza para la mediana de supervivencia dadas las covariables de los sujetos. Sea $\xi_p(\boldsymbol{x})$ que denota el p-ésimo cuantil de la distribución de tiempo de falla para un vector de covariables $\boldsymbol{x}$ y $\widehat{\xi}_p(\boldsymbol{x})$ el p-ésimo cuantil estimado de la función de distribución precedente, su enfoque se basa en la aproximación de normalidad de $\left\{\widehat{\xi}_p(\boldsymbol{x})-\xi_p(\boldsymbol{x})\right\}/\widehat{se}_p(\boldsymbol{x})$, o en su proceso límite gaussiano indexado por $\boldsymbol{x}$, donde $\widehat{se}_p(\boldsymbol{x})$ denota la estimación del error estándar de $\widehat{\xi}_p(\boldsymbol{x})$.
La varianza de la distribución normal límite de $\sqrt{n}\left\{\widehat{\xi}_p(\boldsymbol{x})-\xi_p(\boldsymbol{x})\right\}$ implica la función de riesgo subyacente $\lambda(t)=(d/dt)\,\Lambda(t)$. Aunque Dabrowska y Doksum (1987) citan a Tsiatis (1981) y Andersen y Gill (1982) alegando consistencia sobre su estimador propuesto de varianza límite, Tsiatis, Anderson y Gill únicamente han establecido consistencia para el estimador de Breslow de Λ pero no de la derivada λ. Burr y Doss (1993) hacen uso de la suavización de $\widehat{\Lambda}$ para estimar λ, y en lugar de aplicar la teoría asintótica de $\left\{\widehat{\xi}_p(\boldsymbol{x})-\xi_p(\boldsymbol{x})\right\}/\widehat{se}_p(\boldsymbol{x})$ directamente para construir intervalos de confianza para $\xi_p(\boldsymbol{x})$, la usan para proveer una justificación teórica del método bootstrap-t para construir intervalos de confianza. Sin embargo, como ha sido señalado por Efron y Tibshirani (1993), el método bootstrap-t requiere estimaciones estables de los errores estándar para que funcione bien en la práctica. Por tanto, las dificultades en estimar el error estándar de $\widehat{\xi}_p(\boldsymbol{x})$ también causan dificultades con intervalos de confianza bootstrap-t para $\xi_p(\boldsymbol{x})$.

En efecto, incluso sin censura y efectos covariable de modo que el problema se reduzca a intervalos de confianza para el p-ésimo cuantil ξ_p de una función de distribución basada en una muestra de tiempos de supervivencia independientes e idénticamente distribuidos $t_1, ..., t_n$ con función de densidad común f que tiene un estimador consistente $\widehat{f}$, la distribución normal limite de

$$\widehat{f}(\xi_p)\{n/[p(1-p)]\}^{1/2}\left(\widehat{\xi}_p - \xi_p\right)$$

raramente es usada en la construcción de intervalos de confianza para ξ_p. Además de los problemas con el rendimiento de muestras finitas del estimador de densidad $\widehat{f}\left(\widehat{\xi}_p\right)$, la adecuación de la aproximación lineal $f(\xi_p)\left(\widehat{\xi}_p - \xi_p\right)$ a $F\left(\widehat{\xi}_p\right) - F(\xi_p)$ usada para derivar la normalidad asintótica de $\widehat{\xi}_p - \xi_p$ (donde F es la función de distribución cuya derivada es f) se vuelve complicada cuando $\widehat{\xi}_p$ no es suficientemente cercana a ξ_p. En su lugar, un intervalo de confianza estándar no paramétrico de la forma $t_{(k_1)} < \xi_p < t_{(k_2)}$, donde $t_{(i)}$ denota el estadístico de orden de la muestra y $k_1 < k_2$ son enteros tales que

$$P\left(t_{(k_1)} \leq \xi_p < t_{(k_2)}\right) = P(k_1 \leq B(n,p) < k_2) \geq 1 - 2\alpha.$$

El límite inferior $1-2\alpha$ en (1.4) puede no ser alcanzable debido a la discontinuidad de la distribución binomial $B(n,p)$. Posteriormente Ho y Lee (2005) hicieron uso de iteraciones bootstrap suavizadas para lograr errores de cobertura más precisos de un solo lado del intervalo percentil bootstrap. Este método, sin embargo, es muy intensivo computacionalmente e implica una capa adicional de bootstrapping para determinar el parámetro utilizado para suavizar la distribución empírica.

Para los datos de supervivencia censurados sin covariables, Li et al. (1996) hizo uso de la verosimilitud empírica para construir bandas de confianza para ξ_p, de forma conjunta en $p_1 \leq p \leq p_2$. Sus resultados sobre las probabilidades de cobertura se basan en la convergencia débil. Ellos sin embargo, no tienen suavizada la función de verosimilitud empírica, ni compararon el enfoque de verosimilitud empírica con otros métodos de prueba basados en construir intervalos de confianza para ξ_p cuando los t_i están sujetos a censura. Estos intervalos basados en pruebas estadísticas alternativas se remontan a Brookmeyer y Crowley (1982) quienes invierten una prueba de signos generalizados, que conducen a un conjunto de confianza $1-2\alpha$ aproximado de la forma

$$\left\{t : \left|\widehat{S}(t) - 1/2\right| \leq z_{(1-\alpha)}\widehat{\sigma}(t)\right\}$$

para la mediana $\xi_{1/2}$, donde $\widehat{S}(t)$ es el estimador de Kaplan-Meier de la función de supervivencia, $\widehat{\sigma}(t)$ es el error estándar estimado de $\widehat{S}(t)$ y z_q denota el q-ésimo cuantil de la distribución normal estándar.

En lugar de utilizar la aproximación normal, Strawderman et al. (1997) utiliza expansiones Edgeworth para la función de riesgo acumulada studentizada para obtener límites de confianza basados en pruebas estadísticas más precisas para ξ_p.

En este trabajo se desarrolla un nuevo método para construir intervalos de confianza para el cuantil $\xi_p(\boldsymbol{x})$ en el modelo de riesgos proporcionales (ecuación 1.1). A diferencia de los métodos de Dabrowska y Doksum (1987) y Burr y Doss (1993) que usan $\left\{\widehat{\xi}_p(\boldsymbol{x}) - \xi_p(\boldsymbol{x})\right\}/\widehat{se}_p$ como un pivote aproximado, se usa un enfoque basado en pruebas estadísticas, usando $\widehat{\Lambda}(t \mid \boldsymbol{x})$ para probar si $\Lambda(t \mid \boldsymbol{x}) = \log\left(p^{-1}\right)$, donde $\widehat{\Lambda}(t \mid \boldsymbol{x}) = \widehat{\Lambda}(t)\exp\left(\widehat{\boldsymbol{\beta}}'\boldsymbol{x}\right)$ y $\widehat{\Lambda}(t)$ es el estimador de Breslow de la función de riesgo acumulada subyacente $\Lambda(t)$. En lugar de usar la aproximación normal como en Strawderman et al. (1997) para encontrar los cuantiles de la estadística de prueba, se usa el método bootstrap para evaluar los cuantiles de un pivot aproximado obtenido por studentizar la estadística de prueba.

3.1. Una nueva prueba basada en intervalos de confianza Bootstrap

Una generalización del intervalo de confianza dado por Brookmeyer-Crowley (1.5) para la mediana $\xi_{1/2}(\boldsymbol{x})$ en el modelo de Cox esta dada por

$$\left\{t : \left|\widehat{S}(t \mid \boldsymbol{x}) - (1-p)\right| \leq z_{1-\alpha}\widehat{\sigma}(t \mid \boldsymbol{x})\right\}, \tag{3.1}$$

donde $\widehat{\sigma}^2(t \mid \boldsymbol{x})$ es la varianza asintótica de

$$\widehat{S}(t \mid \boldsymbol{x}) = \exp\left\{-\widehat{\Lambda}(t)\, e^{\widehat{\boldsymbol{\beta}}'\boldsymbol{x}}\right\}, \tag{3.2}$$

en el cual $\widehat{\boldsymbol{\beta}}$ es el estimador que maximiza la ecuación (1.2) y $\widehat{\Lambda}$ es la estimación de Breslow (1974) de la función de riesgo acumulada subyacente basada en $\left(\widetilde{t}_i, \delta_i, x_i\right)$, $1 \leq i \leq n$.
La fórmula de la varianza asintótica fue obtenida por Tsiatis (1981) utilizando el método Delta ($\widehat{\sigma}^2(t \mid \boldsymbol{x}) = (\widehat{S}(t \mid \boldsymbol{x}))^2 v(t \mid \boldsymbol{x})$). Notar que esta varianza asintótica es una función no lineal de la matriz de covarianza asintótica de $(\widehat{\Lambda}(t) - \Lambda(t), (\widehat{\beta} - \beta)'x)$. Aunque $\widehat{S}(t \mid \boldsymbol{x})$ toma valores en el intervalo $[0,1]$, $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{x}$ no tiene dicha limitación y su varianza en muestras finitas puede ser significativa. Por otra parte, la aproximación normal a $\left|\widehat{S}(t \mid \boldsymbol{x}) - S(t \mid \boldsymbol{x})\right|/\widehat{\sigma}(t \mid \boldsymbol{x})$ usada en (3.1) puede ser inadecuada cuando el tamaño de la muestra no es lo suficientemente grande.
En lugar de utilizar $\widehat{S}(t \mid \boldsymbol{x}) - (1-p)$ como la estadística de prueba, se usa la transformación logarítmica para convertirla en $\widehat{\Lambda}(t \mid \boldsymbol{x}) - \log(1-p)^{-1}$. Una ventaja de esta transformación es que a diferencia de $\widehat{S}(t \mid \boldsymbol{x})$, $\widehat{\Lambda}(t \mid \boldsymbol{x})$ ya no está limitada de pertenecer al intervalo $[0,1]$ y por lo tanto la variabilidad debida a $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'x$ en su fórmula de varianza asintótica puede ser compatible con su magnitud. Otra ventaja es que la varianza asintótica de $\widehat{S}(t \mid \boldsymbol{x})$ implica una mayor aproximación lineal alrededor de $\widehat{\Lambda}(t \mid \boldsymbol{x})$. De hecho, tras derivar la varianza asintótica $v(t \mid \boldsymbol{x})$ de:

$$\widehat{\Lambda}(t \mid \boldsymbol{x}) = \widehat{\Lambda}(t)\exp\left(\widehat{\boldsymbol{\beta}}'\boldsymbol{x}\right) \tag{3.3}$$

a partir de la matriz de covarianza asintótica de $(\widehat{\Lambda}(t) - \Lambda(t), (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{x})$, Tsiatis (1981) la

utilizó para derivar la varianza asintótica de $\widehat{S}(t|\boldsymbol{x})$ a través de la transformación no lineal $\widehat{S}(t \mid \boldsymbol{x}) = e^{-\widehat{\Lambda}(t|\boldsymbol{x})}$.
Dado $\boldsymbol{x_i} = (x_{i1}, ..., x_{ik})'$ y $\boldsymbol{a} = (a_1, ..., a_k)'$, se define

$$W(t) = \sum_{j:\widetilde{t}_j \geq t} \exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{x_j}\right), W_l(t) = \sum_{j:\widetilde{t}_j \geq t} x_{jl} \exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{x_j}\right), \tag{3.4}$$

$$Q_l(t,a) = \sum_{i:\widetilde{t}_i \leq t} \delta_i \left\{W_l\left(\widetilde{t}_i\right)/W\left(\widetilde{t}_i\right) - a_l\right\}/W\left(\widetilde{t}_i\right), \tag{3.5}$$

y $\boldsymbol{Q}(t,a) = (Q_1(t,a), ..., Q_k(t,a))'$. Reemplazando los parámetros desconocidos en $v(t \mid \boldsymbol{x})$ por sus estimadores consistentes se obtiene:

$$\widehat{v}(t \mid \boldsymbol{x}) = e^{2\widehat{\boldsymbol{\beta}}' \boldsymbol{x}} \left\{\sum_{i:\widetilde{t}_i \leq t} \delta_i / W^2\left(\widetilde{t}_i\right) + \left(Q(t, \boldsymbol{x})\right)' \left(-\ddot{l}\left(\widehat{\boldsymbol{\beta}}\right)\right)^{-1} Q(t, \boldsymbol{x})\right\}, \tag{3.6}$$

que a su vez origina:

$$\widehat{\sigma}^2(t \mid \boldsymbol{x}) = \left(\widehat{S}(t \mid \boldsymbol{x})\right)^2 v(t \mid \boldsymbol{x}), \tag{3.7}$$

aplicando el método Delta a la transformación $\widehat{S}(t \mid \boldsymbol{x}) = e^{-\widehat{\Lambda}(t|\boldsymbol{x})}$; ver Tsiatis (1981).
En lugar de los cuantiles normales $z_{1-\alpha}$ y $z_\alpha = -z_{1-\alpha}$ usados en (3.1), se aproximan los α-ésimo y $(1-\alpha)$-ésimo cuantiles $c_\alpha(t)$ y $c_{1-\alpha}(t)$ por los cuantiles $\widehat{c}_\alpha(t)$ y $\widehat{c}_{1-\alpha}(t)$ de la distribución bootstrap de $\left\{\widehat{\Lambda}(t \mid \boldsymbol{x}) - \Lambda(t \mid \boldsymbol{x})\right\}/\widehat{v}^{1/2}(t \mid \boldsymbol{x})$. Finalmente, se define la prueba basada en el conjunto de confianza

$$T = \left\{t : \widehat{c}_\alpha(t) \leq \left[\widehat{\Lambda}(t \mid \boldsymbol{x}) - \log(1-p)^{-1}\right]/\widehat{v}^{1/2}(t \mid \boldsymbol{x}) \leq \widehat{c}_{1-\alpha}(t)\right\} \tag{3.8}$$

para el p-ésimo cuantil $\xi_p(\boldsymbol{x})$ dado un vector de covariables $\boldsymbol{x}$.

3.2. Teoría asintótica

Cuando no hay covariables, Lai y Wang (1993) han derivado expansiones Edgeworth para la distribución muestral y también para la distribución bootstrap de $\left\{\widehat{\Lambda}(t) - \Lambda(t)\right\}/\widehat{v}^{1/2}(t)$. En el modelo de regresión de Cox con covariables univariadas, Gu (1992) ha obtenido una ampliación Edgeworth, con error $o\left(n^{-1/2}\right)$, para $Z = \left(-\ddot{l}\left(\widehat{\boldsymbol{\beta}}\right)\right)^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ y también para su contraparte bootstrap $Z^* = \left(-\ddot{l}^*\left(\widehat{\boldsymbol{\beta}}^*\right)\right)^{1/2}\left(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right)$ bajo ciertas condiciones de regularidad ($\widehat{\beta}^*$ es la estimación del vector de parámetros bootstrap); sus argumentos pueden ser fácilmente extendidos a covariables multidimensionales. Su derivación consiste en mostrar que Z y Z^* son U-estadísticas asintóticas (Lai y Wang (1993)). Ya que la estimación de Breslow de la función de riesgo subyacente tiene la forma $\widehat{\Lambda}(t) = \sum_{i:\widetilde{t}_i \leq t}\left\{\delta_i / \sum_{j:\widetilde{t}_j \geq \widetilde{t}_i} \exp\left(\widehat{\boldsymbol{\beta}}' x_j\right)\right\}$, argumentos similares a los de Lai y Wang (1993) pueden ser utilizados para demostrar que

$\widehat{\Lambda}(t) - \Lambda(t)$ es una U-estadística asintótica. A partir de $\widehat{\Lambda}(t \mid \boldsymbol{x}) = e^{\widehat{\boldsymbol{\beta}}' \boldsymbol{x}} \widehat{\Lambda}(t)$, argumentos similares a los de Gu (1992) y a los de Gross y Lai (1996) pueden ser utilizados para probar que $\left\{\widehat{\Lambda}(t \mid \boldsymbol{x}) - \Lambda(t \mid \boldsymbol{x})\right\} / \widehat{v}^{1/2}(t \mid \boldsymbol{x})$ es una U-estadística asintótica que tiene una expansión Edgeworth con error $o\left(n^{-1/2}\right)$.
Como en Gu (1992), se asumen las siguientes condiciones de regularidad:

- (x_i, t_i, c_i) son i.i.d., x_i es acotada (limitada), y t_i y c_i son condicionalmente independientes dado x_i.
- Λ tiene derivada continua λ.
  Además, se asume que $||\boldsymbol{\beta}|| < B$ para algún B conocido y que
- $P\left(\widetilde{t}_i \geq \tau\right) > 0$ para algún $\tau > \xi_p(\boldsymbol{x})$ conocido, y se redefine la ecuación (1.2) por:

$$l(\boldsymbol{\beta}) = \sum_{i:\widetilde{t}_i \geq \tau} \delta_i \left\{ \beta' x_i - \log \left(\sum_{j:\widetilde{t}_j \geq \widetilde{t}_i} \exp\left(\beta' x_j\right) \right) \right\},$$

  de modo que $\widehat{\boldsymbol{\beta}}$ es el estimador que maximiza esta modificación de (1.2) dentro del conjunto acotado $\{\beta : ||\beta|| \leq B\}$. Dado que $\tau > \xi_p(\boldsymbol{x})$, se puede también modificar el conjunto de confianza T, definido por la ecuación (3.8), al restringir $\{t : t \leq \tau\}$ de modo que los resultados de Lai y Wang (1993) en U-estadísticas asintóticas y expansiones Edgeworth pueden ser aplicadas a $\widehat{\Lambda}(t) - \Lambda(t)$ para cada $t \in T$. Además, se asume que:
- $\int_0^\tau \left\{\alpha_2(t) - \alpha_1(t)\, \alpha_1{}'(t) / \alpha_0(t)\right\} \lambda(t)\, dt$ es una matriz definida positiva,
  donde $\alpha_k(t) = E\left(\boldsymbol{x}^k e^{\boldsymbol{\beta}' \boldsymbol{x}} I_{\{\widetilde{t} \geq t\}}\right)$ para $k = 0, 1, 2$, con $x^0 = 1$ y $x^2 = xx'$.

Bajo estos supuestos, no sólo $\left\{\widehat{\Lambda}(t \mid \boldsymbol{x}) - \Lambda(t \mid \boldsymbol{x})\right\} / \widehat{v}^{1/2}(t \mid \boldsymbol{x})$ tiene una expansión Edgeworth con error $o\left(n^{-1/2}\right)$, los coeficientes de esta expansión Edgeworth también difieren de la contraparte bootstrap $\left\{\widehat{\Lambda}^*(t \mid \boldsymbol{x}) - \Lambda(t \mid \boldsymbol{x})\right\} / \widehat{v}^{*1/2}(t \mid \boldsymbol{x})$ por $o_p\left(n^{-1/2}\right)$, ver (Gu (1992)). Por lo tanto,

$$\widehat{c}_\alpha(t) - c_\alpha(t) = o_p\left(n^{-1/2}\right), \widehat{c}_{1-\alpha}(t) - c_{1-\alpha}(t) = o_p\left(n^{-1/2}\right) \tag{3.9}$$

para cada t fijo. Aplicando la ecuación (3.9) y argumentos similares al de Hall (1992), entonces los rendimientos de la ecuación (3.8) son:

$$\begin{aligned} P(\xi_p(\boldsymbol{x}) \in T) &= P\left(\widehat{c}_\alpha(\xi_p(\boldsymbol{x})) \leq \left[\widehat{\Lambda}(\xi_p(\boldsymbol{x}) \mid \boldsymbol{x}) - \Lambda(\xi_p(\boldsymbol{x}) \mid \boldsymbol{x})\right] / \widehat{v}^{1/2}(\xi_p(\boldsymbol{x}) \mid \boldsymbol{x}) \leq \widehat{c}_{1-\alpha}(\xi_p(\boldsymbol{x}))\right) \\ &= 1 - 2\alpha + o\left(n^{-1/2}\right). \end{aligned} \tag{3.10}$$

3.3. Cálculo de los límites de confianza

El conjunto dado en la ecuación (3.8) no puede ser un intervalo, como fue señalado por Brookmeyer y Crowley (1982) para el conjunto de confianza dado en (1.5) cuando no hay covariables. En la práctica, a menudo basta con dar sólo el límite superior e inferior de (3.8), obteniéndose de este modo un intervalo de confianza. Sea $q = \alpha$ ó $1-\alpha$, notar que para un $\boldsymbol{x}$ fijo, la función de riesgo acumulada Λ es una función escalonada con saltos ($\delta_i = 1$) en las observaciones no censuradas $\widetilde{t}_i$, y por tanto es la función $\widehat{v}$. Los saltos en las $\widehat{t}_i's$ no censuradas también causan discontinuidades de $\widehat{c}_q$ en estos puntos. Sea $\left[\widetilde{\Lambda}(. \mid \boldsymbol{x}) - \log(1-p)^{-1}\right] / \widetilde{v}^{1/2}(. \mid \boldsymbol{x}) - \widetilde{c}_q(.)$ que denota la modificación de $\left[\widehat{\Lambda}(. \mid \boldsymbol{x}) - \log(1-p)^{-1}\right] / \widehat{v}^{1/2}(. \mid \boldsymbol{x}) - \widehat{c}_q(.)$ que interpola linealmente entre los valores correspondientes de dos $\widehat{t}_i's$ no censurados adyacentes.

Se asume que las covariables x_i son i.i.d., entonces la distribución bootstrap del pivote asintótico $\left\{\widehat{\Lambda}(t \mid \boldsymbol{x}) - \Lambda(t \mid \boldsymbol{x})\right\} / \left(\widehat{v}(t \mid \boldsymbol{x})\right)^{1/2}$ puede ser evaluada por remuestreo de $\left\{\left(\widetilde{t}_i, \delta_i, x_i\right) : 1 \leq i \leq n\right\}$ para obtener B muestras bootstrap $\left\{\left(\widetilde{t}_i^*, \delta_i^*, x_i^*\right)_b, 1 \leq i \leq n\right\}, 1 \leq b \leq B$. En cada valor dado de t, se calcula $w_b^*(t) = \left\{\widehat{\Lambda}_b^*(t \mid \boldsymbol{x}) - \Lambda(t \mid \boldsymbol{x})\right\} / \left(\widehat{v}_b^*(t \mid \boldsymbol{x})\right)^{1/2}$ a partir de la b-ésima muestra bootstrap, y los α-ésimo y $(1-\alpha)$-ésimo cuantiles de $\{w_1^*, ..., w_B^*\}$ se calculan para obtener $\widehat{c}_\alpha(t)$ y $\widehat{c}_{1-\alpha}(t)$. Se puede utilizar el siguiente procedimiento iterativo para elegir los valores de t, pertenecientes al conjunto ordenado U de $\widehat{t}_i's$ no censurados, en el que $\widehat{c}_\alpha(t)$ o $\widehat{c}_{1-\alpha}(t)$ son calculados. Para terminar, se considera $\widehat{c}_\alpha(t)$. El objetivo del proceso iterativo es resolver la ecuación $g(t) = 0$, donde:

$$g(t) = \left\{\widetilde{\Lambda}(t \mid \boldsymbol{x}) - \log(1-p)^{-1}\right\} / \widetilde{v}^{1/2}(t \mid \boldsymbol{x}) - \widetilde{c}_\alpha(t). \tag{3.11}$$

Sean a y b los elementos más pequeño y más grande de U respectivamente. Con $g(a) < 0$ y $g(b) > 0$, se puede utilizar el método de bisección para encontrar dos elementos adyacentes de U donde g cambia de signo. Entonces se interpola linealmente entre estos dos puntos para encontrar la solución de $g(t) = 0$, o simplemente se toma el elemento más grande para ser el límite de confianza. Notar que este procedimiento también se puede utilizar para calcular pruebas basadas en intervalos de confianza bootstrap para los cuantiles ξ_p en ausencia de covariables y también en el caso de observaciones i.i.d. completas.

Burr y Doss (1993) utilizan otro sistema de remuestreo bajo el supuesto de que las variables con censura c_i tienen la misma función de distribución de C. Sea $\widehat{C}$ el estimador Kaplan-Meier de C. Una muestra bootstrap es de la forma $\left\{\left(\widetilde{t}_i^*, \delta_i{}^*, x_i\right) : 1 \leq i \leq n\right\}$, donde $\widetilde{t}_i^* = \text{mín}\left(t_i^*, c_i^*\right)$ y $\delta_i^* = I_{\left\{t_i^* \leq c_i^*\right\}}$, en el que $c_i{}^*$ es generado de $\widehat{C}$ y t_i^* es generado de $\widehat{S}(. \mid x_i)$ independientemente de c_i^*. Este esquema de remuestreo no necesita que los x_i sean idénticamente distribuidos pero asume que los c_i sean idénticamente distribuidos en su lugar.

Capítulo 4

Implementación computacional

En esta sección se exponen las fórmulas y los diagramas de flujo asociados a las pruebas estadísticas empleadas para la obtención de los intervalos.

4.1. Para la construcción de los intervalos de confianza

Para proceder con el cálculo de los intervalos de confianza se citan a continuación las ecuaciones que permiten realizar su estimación.

Así se tiene que una generalización del intervalo de confianza para la mediana $\xi_{1/2}(x)$ en el modelo de Cox está dada por la ecuación (3.1):

$$\left\{t : \left|\widehat{S}(t \mid x) - (1-p)\right| \leq z_{1-\alpha}\widehat{\sigma}(t \mid x)\right\},$$

donde $\widehat{\sigma}^2(t \mid x)$ es la varianza asintótica de

$$\widehat{S}(t \mid x) = \exp\left\{-\widehat{\Lambda}(t)\, e^{\widehat{\beta}' x}\right\},$$

expresada en la ecuación (3.2), en la cual $\widehat{\beta}$ es el estimador de máxima verosimilitud y $\widehat{\Lambda}$ es el estimador de Breslow de la función de riesgo acumulada subyacente basada en $(\widetilde{t}_i, \delta_i, x_i)$, $1 \leq i \leq n$.

Recordamos que la estimación de Breslow de la función de riesgo subyacente tiene la forma:

$$\widehat{\Lambda}(t) = \sum_{i:\widetilde{t}_i \leq t} \left\{ \delta_i / \sum_{j:\widetilde{t}_j \geq \widetilde{t}_i} \exp\left(\widehat{\boldsymbol{\beta}}' x_j\right) \right\} \tag{4.1}$$

y la función de riesgo acumulada, la expresada en la ecuación (3.3):

$$\widehat{\Lambda}(t \mid \boldsymbol{x}) = e^{\widehat{\boldsymbol{\beta}}' \boldsymbol{x}} \widehat{\Lambda}(t)$$

Dado $x_i = (x_{i1}, ..., x_{ik})'$ y $a = (a_1, ..., a_k)'$, se define

$$W(t) = \sum_{j:\widetilde{t}_j \geq t} \exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{x_j}\right), W_l(t) = \sum_{j:\widetilde{t}_j \geq t} x_{jl} \exp\left(\widehat{\boldsymbol{\beta}}' \boldsymbol{x_j}\right),$$
$$Q_l(t,a) = \sum_{i:\widetilde{t}_i \leq t} \delta_i \left\{W_l\left(\widetilde{t}_i\right)/W\left(\widetilde{t}_i\right) - a_l\right\}/W\left(\widetilde{t}_i\right),$$

y $Q(t,a) = (Q_1(t,a), ..., Q_k(t,a))'$. Reemplazando los parámetros desconocidos en $v(t \mid x)$ por sus estimadores consistentes se obtiene la ecuación (3.6):

$$\widehat{v}(t \mid x) = e^{2\widehat{\beta}' x}\left\{\sum_{i:\widetilde{t}_i \leq t} \delta_i / W^2\left(\widetilde{t}_i\right) + (Q(t,x))'\left(-\ddot{l}\left(\widehat{\beta}\right)\right)^{-1} Q(t,x)\right\},$$

que a su vez origina la ecuación (3.7):

$$\widehat{\delta}^2(t \mid x) = \left(\widehat{S}(t \mid x)\right)^2 v(t \mid x),$$

aplicando el método Delta a la transformación $\widehat{S}(t \mid x) = e^{-\widehat{\Lambda}(t|x)}$.
En lugar de los cuantiles normales $z_{1-\alpha}$ y $z_\alpha = -z_{1-\alpha}$ usados en la ecuación (3.1), se aproximan los α-ésimo y $(1-\alpha)$-ésimo cuantiles $c_\alpha(t)$ y $c_{1-\alpha}(t)$ por los cuantiles $\widehat{c}_\alpha(t)$ y $\widehat{c}_{1-\alpha}(t)$ de la distribución bootstrap de $\left\{\widehat{\Lambda}(t \mid x) - \Lambda(t \mid x)\right\}/\widehat{v}^{1/2}(t \mid x)$. Finalmente, se define la prueba basada en el conjunto de confianza dado en la ecuación (3.8):

$$T = \left\{t : \widehat{c}_\alpha(t) \leq \left[\widehat{\Lambda}(t \mid x) - \log(1-p)^{-1}\right]/\widehat{v}^{1/2}(t \mid x) \leq \widehat{c}_{1-\alpha}(t)\right\}$$

para el p-ésimo cuantil $\xi_p(x)$ dado un vector de covariables x.

El diagrama de flujo para la construcción de los intervalos de confianza esta dado en las Figuras 4.1, 4.2 y 4.3.

4.2. Para el cálculo de los cuantiles Bootstrap

El procedimiento que permite obtener los límites de confianza parte de asumir que las covariables x_i sean i.i.d. con lo que la distribución bootstrap del pivote asintótico $\left\{\widehat{\Lambda}(t \mid x) - \Lambda(t \mid x)\right\}/\left(\widehat{v}(t \mid x)\right)^{1/2}$ puede ser evaluada por remuestreo de $\left\{\left(\widetilde{t}_i, \delta_i, x_i\right) : 1 \leq i \leq n\right\}$ para obtener B muestras bootstrap $\left\{\left(\widetilde{t}_i^*, \delta_i^*, x_i^*\right)_b, 1 \leq i \leq n\right\}, 1 \leq b \leq B$. En cada valor de t, se calcula $w_b^*(t) := \left\{\widehat{\Lambda}_b^*(t \mid x) - \Lambda(t \mid x)\right\}/\left(\widehat{v}_b^*(t \mid x)\right)^{1/2}$ a partir de la b-ésima muestra bootstrap, y se calculan los α-ésimo y $(1-\alpha)$-ésimo cuantiles de $\{w_1^*, ..., w_B^*\}$ para obtener $\widehat{c}_\alpha(t)$ y $\widehat{c}_{1-\alpha}(t)$. Estos cuantiles se reemplazan en la ecuación (4.1) con lo que se obtienen los intervalos de confianza.

El diagrama de flujo asociado a este procedimiento esta dado en la Figura 4.4.

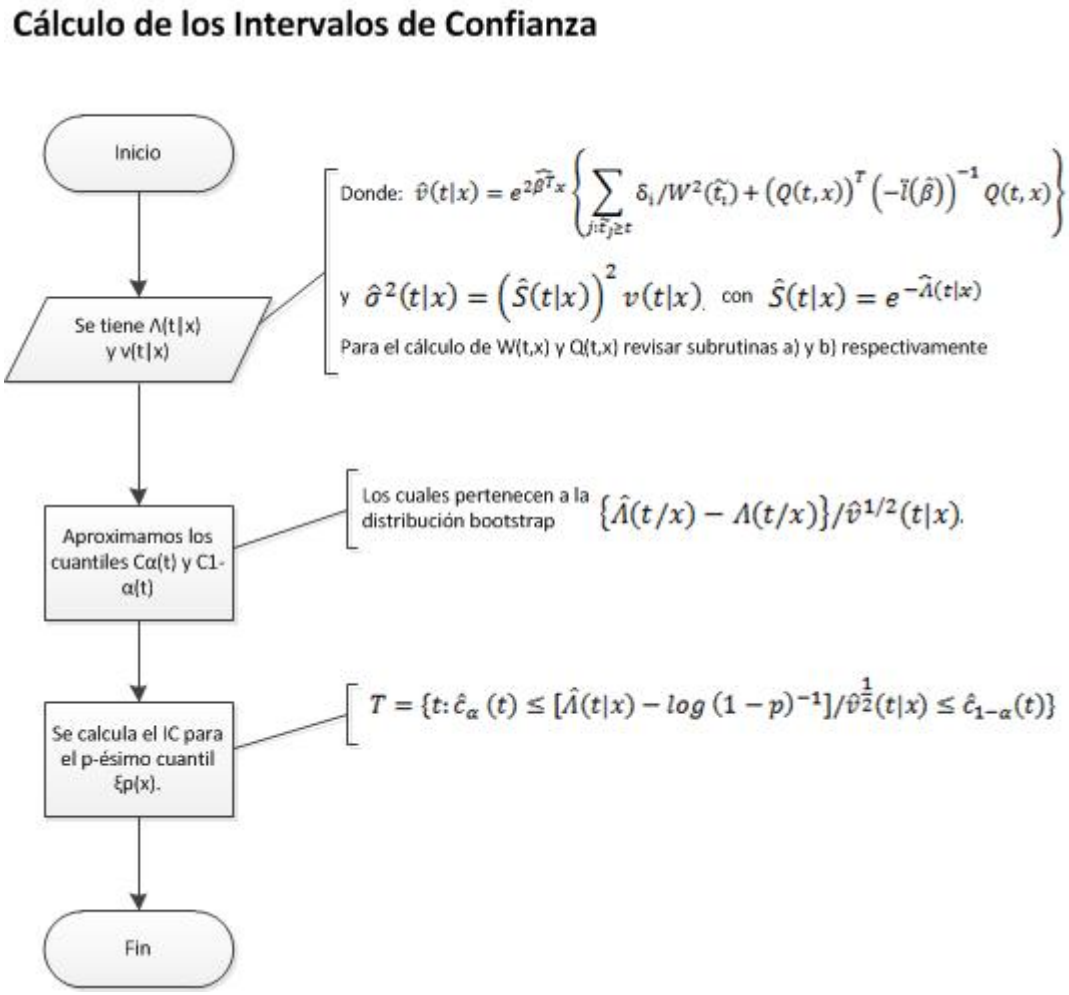

Figura 4.1: Diagrama de flujo para el cálculo de los intervalos de confianza

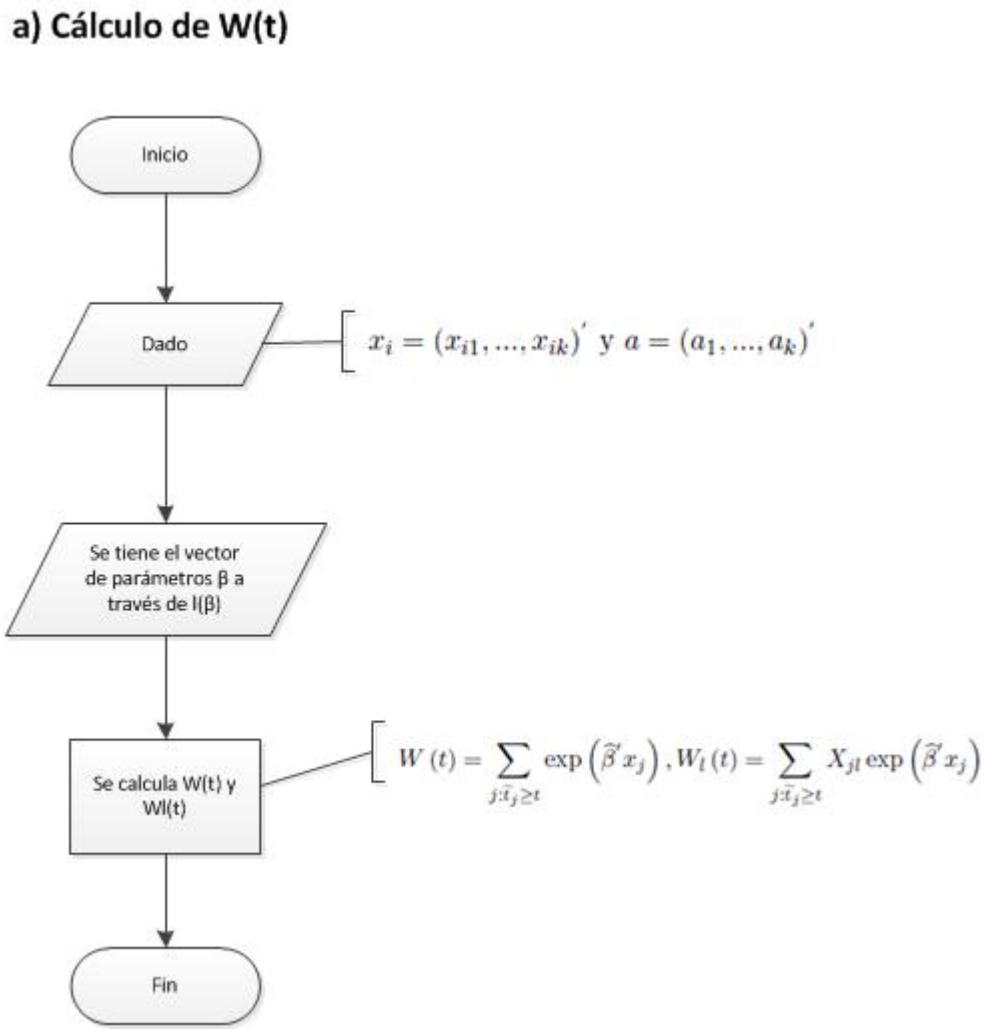

Figura 4.2: Cálculo de los intervalos de confianza - Subrutinas complementarias: Cálculo de W(t) y $W_l(t)$

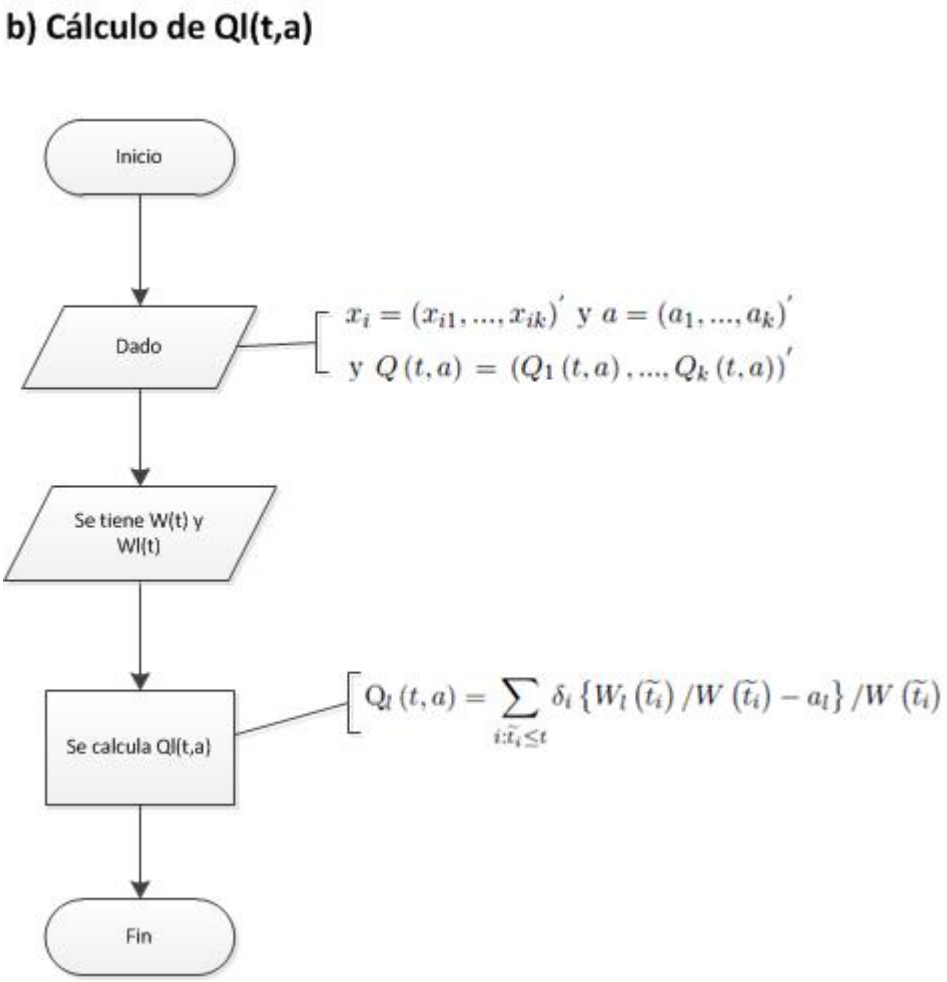

Figura 4.3: Cálculo de los intervalos de confianza - Subrutinas complementarias: Cálculo de $Q_l(t,a)$

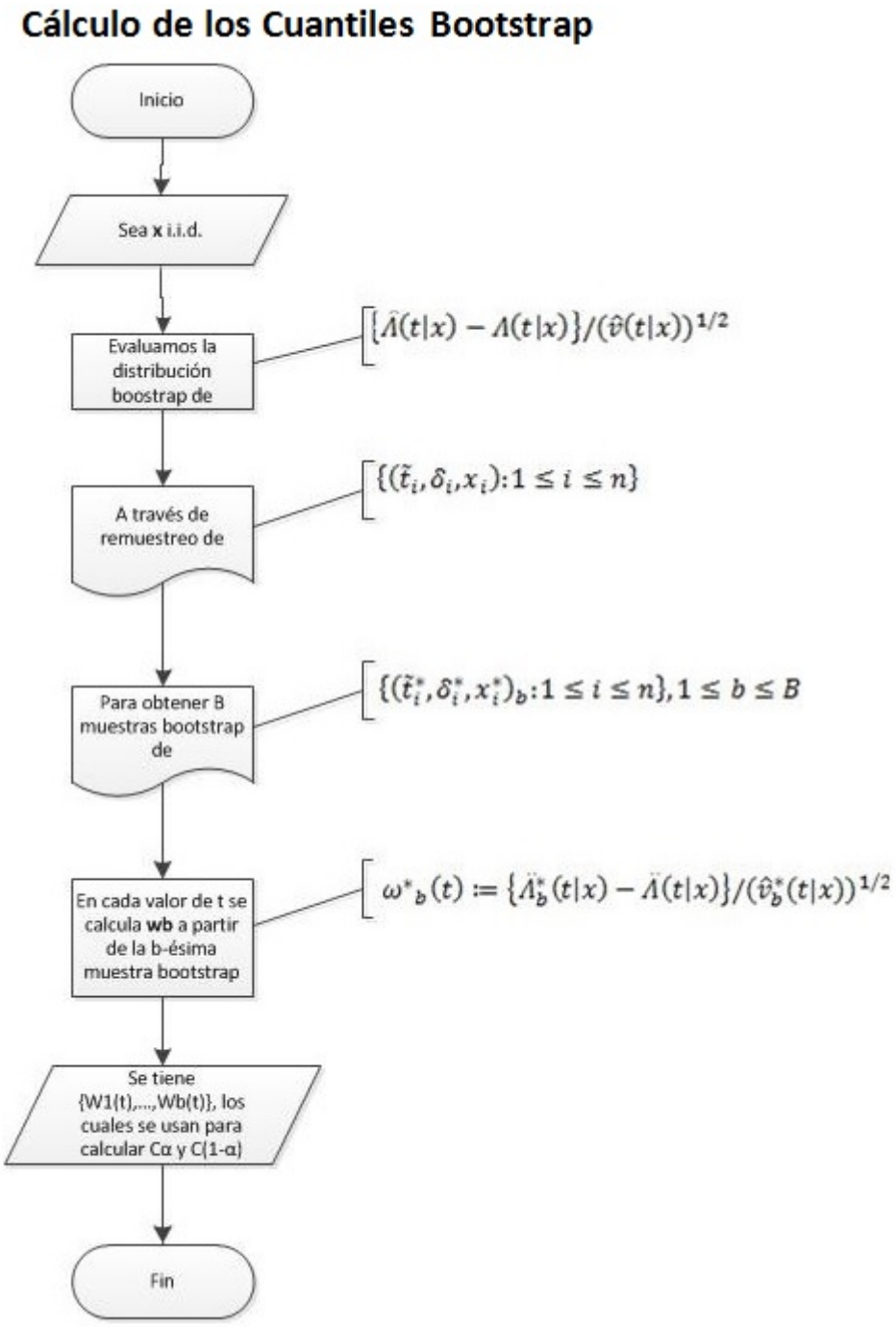

Figura 4.4: Diagrama de flujo para el cálculo de los cuantiles bootstrap.

Capítulo 5

Aplicación

Se realizará la aplicación a un conjunto de datos perteneciente a una empresa de telecomunicaciones [2].
La unidad en análisis es la línea móvil. La aplicación se dará sobre un conjunto de líneas móviles postpago pertenecientes a una campaña de migración.
Mediante la obtención de intervalos de confianza para la mediana de supervivencia de las líneas se conocerá su riesgo de vida promedio así de cómo las covariables inciden sobre el tiempo hasta el incumplimiento del pago de los clientes. Ésto permitirá a los tomadores de decisión de la empresa a realizar acciones con la finalidad de prolongar el tiempo de vida de las líneas.

5.1. Descripción del conjunto de datos

La unidad en análisis es la línea telefónica, unidad elemental para las empresas de telecomunicaciones.
La empresa cuenta con dos tipos, la móvil, compuesta por líneas telefónicas de celulares y de módems. La otra es la línea fija, compuesta por líneas telefónicas fijas.
La empresa agrupa a sus clientes en tres grupos de acuerdo al tipo de línea con la que cuentan. Así se tiene: Prepago, Postpago y Corporativo.
El conjunto de datos en estudio se tomará del segmento Postpago el cual está compuesto por abonados que pagan de manera mensual un cargo fijo por el servicio. Dentro de este grupo, el estudio se realizará sobre las líneas provenientes de la migración de líneas de Prepago a Postpago. Estas migraciones provienen de una campaña de telemarketing, la cual es de gran importancia para la empresa, puesto que éstas representan aproximadamente el 20 % de la base de clientes Postpago.
Las condiciones para seleccionar el conjunto de líneas fueron:

1. Periodo de Migración (de la plataforma Prepago a Postpago): Se tomó el mes de Noviembre del año 2010. Contiene la información de 15, 358 líneas.

2. Periodo de Evaluación: Se estudió el comportamiento de las líneas durante un año.

Se seleccionó un mes de migración debido a que cuenta con la cantidad de líneas suficiente para obtener resultados válidos.

[2] Por fines de confidencialidad de la información no se menciona el nombre de la empresa de telecomunicaciones en referencia.

Se escogió estudiar el comportamiento de las líneas durante doce meses puesto que, por experiencia, se sabe que es el tiempo necesario para la madurez de los clientes con una línea Postpago. Transcurrido dicho tiempo, conocen la mayoría de características de una línea Postpago (formas de pago, facturación, tarifas, promociones) y su diferencia con las de Prepago (montos de recarga, activación de promociones).
Las covariables con las que se dispone son muchas. Luego de evaluar diferentes covariables, las seleccionadas fueron dos, facturación promedio y edad, las cuales influyen de manera significativa en el tiempo de vida de la línea.

Inicialmente los algoritmos se trabajaron con toda la población sin embargo el tiempo computacional requerido para una sola muestra bootstrap, en una computadora con 2 núcleos y 4 gigabytes de RAM, es bastante grande (más de 20 horas). Tomando en cuenta que este proceso se tiene que repetir B veces, se buscó apoyo en la Dirección de Informática de la Universidad (DIRINFO - equipo Legión). Las computadoras que se emplearon contaban con una mayor capacidad: mayor número de núcleos e igual memoria RAM.
El objetivo era correr los algoritmos en varias computadoras de manera paralela. A pesar de ello, si bien es cierto el tiempo obtenido era menor al alcanzado en la computadora con menor número de núcleos, éste aún era significativamente grande (más de 16 horas). Por ejemplo, para el cálculo del algoritmo $Q_l(t, a)$, el tiempo de cálculo era de aprox 2 horas. Se encontró así que el tiempo que demanda la ejecución de cada rutina sigue un comportamiento no lineal.

Debido a lo anterior, se decidió tomar una muestra representativa de la población en estudio. Ya que, producto del análisis exploratorio se aprecia que se forman estratos en torno a la variable factura, se eligió la muestra empleando muestreo estratificado (Cochran (1981)). Al comparar los histogramas de ambas covariables tanto para la muestra como para los de la población se aprecia un comportamiento similar (ver Figura 5.1).
Considerando un nivel de confianza del 95 % y un error del 5 % se obtiene como muestra requerida, 157 líneas. Empleando un error del 2 %, el valor que se necesita es de 932 líneas. En base a éstos resultados, se decidió ampliar el valor de la muestra al 10 % de la población es decir a 1, 535 líneas.

Es así que se procedió a aplicar los algoritmos para la muestra observando que el tiempo requerido es computacionalmente manejable. El tiempo requerido para una sola muestra bootstrap empleando dicha cantidad de datos es de poco más de 20 minutos.

Las variables que se incluyen en el archivo de análisis son:
TIEMPO: Días de actividad (vida) de la línea a través del periodo en estudio.
ESTADO: Condición de censura (1 denota la desactivación de la línea (no censura) y 0 denota los datos censurados).
FACTURA: Cargo fijo mensual (soles con I.G.V.) al momento de la migración de Prepago a Postpago.

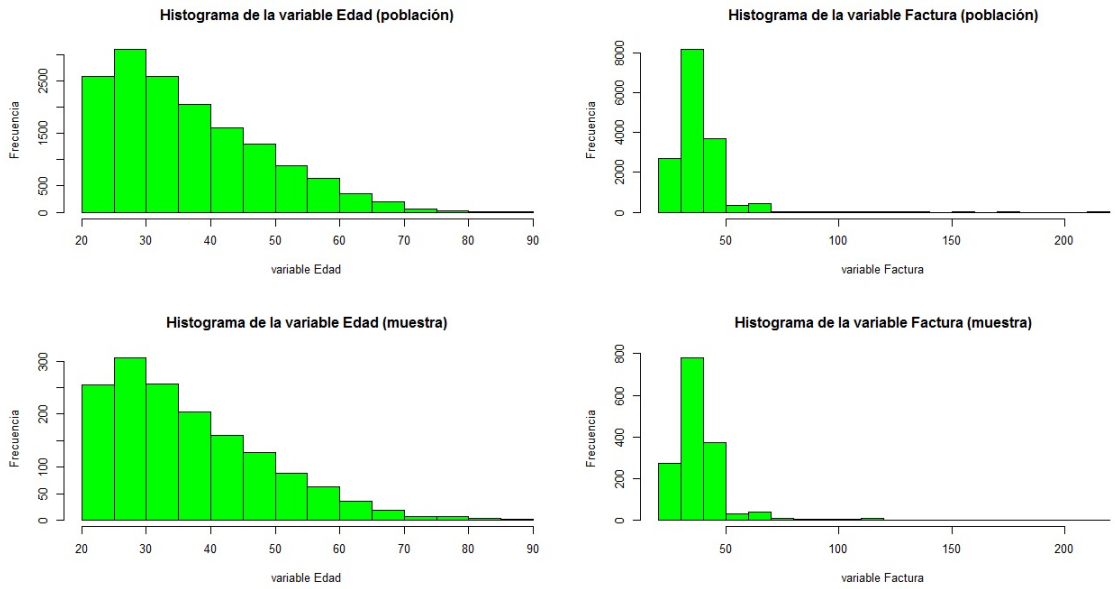

Figura 5.1: Histograma de las covariables edad y factura para la población y la muestra

EDAD: Edad del cliente al inicio del análisis.

Se tienen las estadísticas descriptivas de la muestra en la Figura 5.2.

	n	Media	Error típico	Mediana	Moda	sd	Var	Curtosis	Asimetría	Rango	Min	Max
TIEMPO	1,535	239.76	3.63	238	377	142.29	2.02E+04	-1.68	-0.16	391	3	394
ESTADO	1,535	0.59	0.01	1	1	0.49	0.24	-1.88	-0.35	1	0	1
FACTURA	1,535	39.61	0.35	37	35	13.64	185.93	43.25	5.11	200	20	220
EDAD	1,535	37.05	0.30	34	25	11.94	142.58	0.58	0.95	65	21	86

Figura 5.2: Tabla resumen de estadísticas descriptivas de las covariables del modelo para la muestra

Respecto a la variable edad se puede observar que la media es de 37 años, con un 5 % de clientes con un valor por encima de los 60 años. La forma de la curva parece ser un poco apuntada con una asimetría positiva (ver Figura 5.3).
Respecto a la variable factura se puede observar que el cargo fijo promedio de las lineas migradas es de casi S/.40. La variabilidad de los datos es mayor que el de la variable edad. La forma de la curva es notablemente apuntada (leptocúrtica) presentando también una aparente asimetría positiva (ver Figura 5.4).

5.2. Resultados numéricos

En esta sección se presentan los cálculos realizados para obtener el valor de los parámetros del modelo de Cox así como los resultados obtenidos al aplicar las fórmulas dadas en los diagramas de flujo.

Para obtener el valor de los parámetros del modelo se empleó el software Mathematica (2012). Estos valores fueron corroborados con los software R (2012) y SPSS (2009). Por ejemplo para

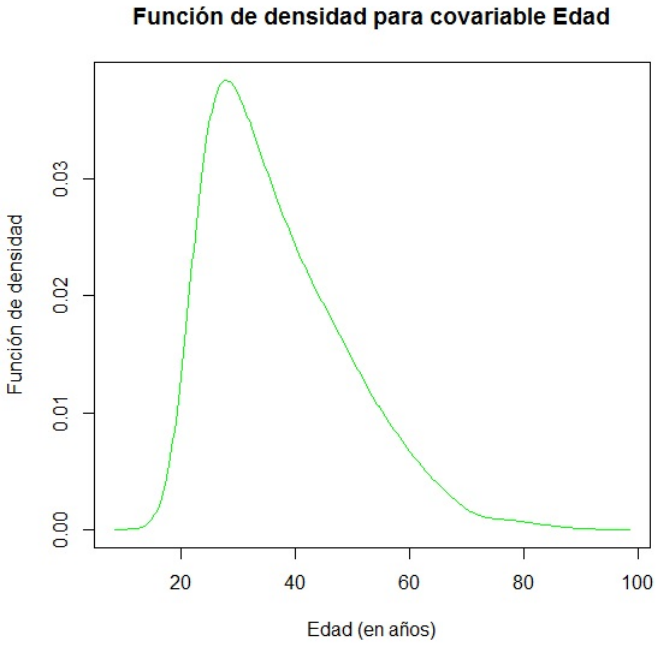

Figura 5.3: Función de densidad para la covariable edad

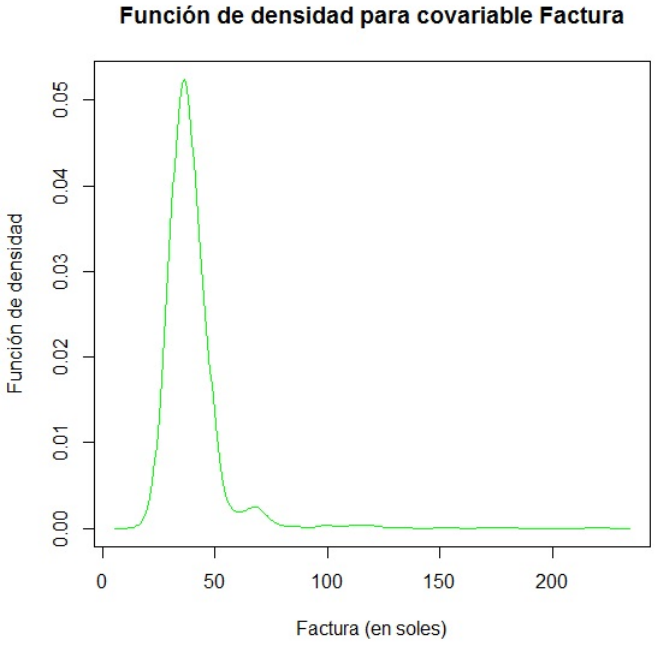

Figura 5.4: Función de densidad para la covariable factura

el software *R*, se emplearon las funciones *coxph* y *Surv* de la librería Survival (2013). Los valores obtenidos se muestran en la Figura 5.5.

Puede afirmarse que las covariables factura y edad son significativas al 5 %, debido a que los p-valores obtenidos son todos menores que 0.05.

Covariables	beta	exp(beta)	exp(-coef)	p-valor
EDAD	-0.006216	0.993804	0.993803	3.19E-02
FACTURA	0.011627	1.011695	1.011695	1.88E-14

Figura 5.5: Modelo de Cox para las líneas móviles y desactivación como evento de interés

De igual forma este modelo resulta significativo por cualquiera de los tres criterios (test de Razón de Verosimilitud, test de Wald, test de Puntajes) para un 5 % de significación, debido a que los p-valores son todos menores que 0.05. Para el test de Razón de Verosimilitud se obtuvo un p-valor de 0.01376, para el test de Wald de 0.01575 y para el test de Puntajes de 0.01569.

Otra información importante, obtenida a través de la salida anterior, es la estimación de los riesgos relativos (a partir de los exp(coef)). En cuanto a la covariable edad, un cliente con una determinada cantidad de años tiene 0.9938 más veces el riesgo de desactivarse en relación a una persona con un año menor. Por otro lado, al aumentar la factura en una unidad monetaria, el riesgo se hace 1.0117 más veces que la del menor valor.

A continuación se presentan los resultados obtenidos al aplicar las fórmulas dadas en los diagramas de flujo.

Se empieza por la lectura de los datos para asignar a cada variable los registros que correspondan.
Se calcula el valor de los parámetros de β y éstas estimaciones se asignan a una variable para que sean posteriormente usados.
Se genera la terna formada por $\left(\widetilde{t}_i, \delta_i, x_i\right)$ como se detalla en la sección (1.1).
Luego se define el periodo de evaluación. Este tiempo se ha definido como una partición de doce meses de treinta días cada una, de esta manera en el análisis se esta considerando a la variable tiempo de manera discreta. Así se tiene entonces un vector de la forma $\{30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330, 365\}$.

Previo al desarrollo de los diagramas de flujo en referencia, se calcula la función de azar subyacente, la función de riesgo acumulada y la de la función de supervivencia ($\widehat{\Lambda}$, $\widehat{\Lambda}\left(t \mid x\right)$ y $\widehat{S}\left(t \mid x\right)$ respectivamente) en base a las ecuaciones expuestas en la sección (4.1). Por ejemplo en la Figura 5.6 se muestra el cálculo de la función de supervivencia para una muestra de diez individuos en los diferentes $t = 12$ tiempos, donde se aprecia que a medida que el tiempo transcurre el valor de la función de supervivencia va decreciendo.
Estas estimaciones son importantes debido a que se van a emplear para cálculos posteriores.

Ahora se aborda el cálculo de $W(t)$, $W_l(t)$ y $Q_l(t, a)$, los dos primeros expuestos en el diagrama de flujo 4.2 y el otro perteneciente al 4.3.
Considerar que para el cálculo de $Q_l(t, a)$ se define previamente el vector a_l, el cual se en-

n	1	2	3	4	5	6	7	8	9	10
t										
1	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
2	0.88	0.88	0.87	0.90	0.89	0.88	0.88	0.89	0.87	0.88
3	0.83	0.82	0.81	0.85	0.84	0.83	0.83	0.84	0.82	0.82
4	0.71	0.70	0.68	0.74	0.72	0.72	0.71	0.72	0.70	0.70
5	0.66	0.65	0.63	0.70	0.68	0.67	0.66	0.68	0.65	0.65
6	0.57	0.56	0.54	0.62	0.59	0.58	0.57	0.59	0.56	0.56
7	0.54	0.53	0.51	0.59	0.56	0.55	0.54	0.56	0.52	0.53
8	0.50	0.49	0.47	0.55	0.52	0.51	0.50	0.52	0.49	0.49
9	0.48	0.48	0.45	0.54	0.50	0.49	0.48	0.51	0.47	0.47
10	0.47	0.46	0.43	0.52	0.49	0.48	0.46	0.49	0.45	0.46
11	0.46	0.45	0.42	0.51	0.48	0.47	0.45	0.48	0.44	0.45
12	0.42	0.41	0.38	0.48	0.44	0.43	0.42	0.44	0.40	0.41

Figura 5.6: Cálculo de la función de supervivencia para una muestra de diez individuos en diferentes $t = 12$ tiempos

cuentra en función de cada covariable. Así pues el valor de $Q_l(t, a)$ sera una función que depende de ambas covariables.

Para obtener $\widehat{v}\,(t \mid x)$ es necesario el cálculo de la inversa de la segunda derivada de la función de máxima verosimilitud del modelo de Cox $\left(\ddot{l}\left(\widehat{\beta}\right)\right)$ evaluada en los parámetros estimados. Para calcular $\ddot{l}\left(\widehat{\beta}\right)$ se emplea el desarrollo realizado en la sección (A.1) del apéndice A. De esta manera $\widehat{v}\,(t \mid x)$ al igual que $Q_l(t, a)$ queda expresado en función de cada covariable. Finalmente se calcula la varianza asintótica $\widehat{\delta}^2\,(t \mid x)$.

Ahora, para realizar el cálculo de los cuantiles bootstrap se requiere el cálculo de $w_b^*\,(t)$. Así se procede a calcular $\widehat{\Lambda}_b^*\,(t \mid x)$ y $\widehat{v}_b^*\,(t \mid x)$ los cuales son producto de realizar B muestreos bootstrap. Con lo anterior se forma el vector $\{w_1^*, ..., w_B^*\}$.
Para realizar el remuestreo vía bootstrap se emplea en el software *Mathematica* la función *RandomChoice*. Como se expuso en la segunda y tercera condición de regularidad de la sección (3.2) y dada la gran cantidad de datos con los que se cuenta, un número de $B = 100$ muestras bootstrap es sugerida para una buena precisión. Un número grande de muestreos bootstrap es requerido cuando se tienen pocos datos (tratamientos clínicos por ejemplo).

Una vez obtenidos los $w_b^*\,(t)$ se procede a calcular los cuantiles 25 % y 75 % (ver diagrama de flujo 4.4). Luego de precisar lo anterior, se procede con el cálculo del conjunto de confianza T para $\widehat{\Lambda}\,(t \mid x)$ (ver diagrama de flujo 4.1).

En la Figura 5.7 se muestra para una línea en general cual es la magnitud de su función de riesgo en base al análisis realizado con las covariables facturación promedio y edad. Así tenemos que para cuando inicia el análisis, es decir en el mes de realizada la migración a postpago, son pocas las líneas que se desactivarán sin embargo conforme transcurre el tiempo esta magnitud se va incrementando. Por ejemplo luego de 6 meses, ésta magnitud llega a un valor de aprox 0.5. Al término del estudio ($t = 360$ días), su valor oscila entre 0.82 y 0.87.
Otro punto que se aprecia es que, conforme avanza el tiempo, la amplitud del intervalo se va acrecentando acentuándose en la parte final del periodo de evaluación. Esto puede deberse a

la pérdida de información en los extremos.

Estos valores guardan relación con la realidad. Se sabe que en los primeros meses de transcurrida la migración, pocas son las líneas que se dan de baja o que migran a prepago. Conforme transcurre el tiempo, por diferentes motivos como por ejemplo el deseo de adquirir nuevos equipos más tecnológicos o el desconocimiento de la renovación de estos terminales, la cantidad de clientes que dan de baja a la línea se tiende a incrementar. De acuerdo a estos resultados, al parecer existe una importante cantidad de clientes que no valoran el conservar su número telefónico o que no están dispuestos a mantener un pago fijo mensual por el servicio.

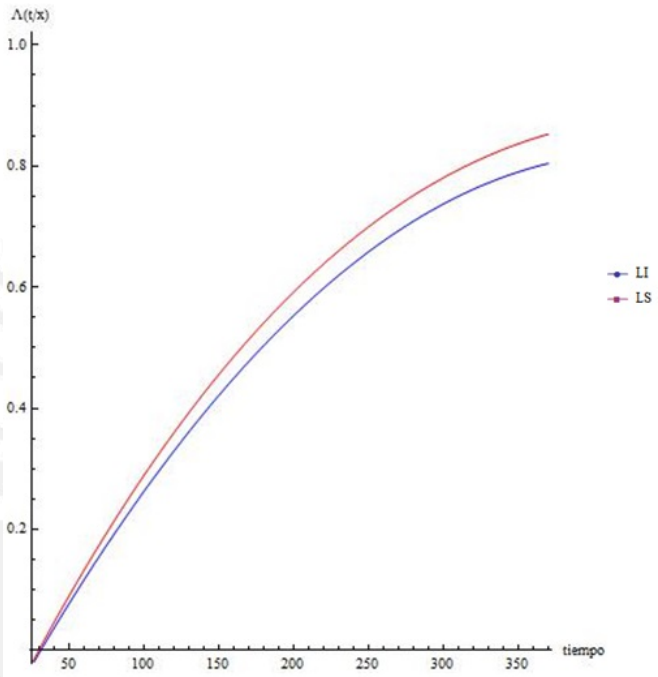

Figura 5.7: Intervalo de confianza para una línea en general a través del tiempo (ajuste obtenido de los pares)

Todo el desarrollo hasta aquí expuesto se encuentra en la sección (B.1) del apéndice B.

Finalmente, como aporte adicional, lo que se realizó fue el considerar a las dos covariables empleadas (factura y edad) como libres, de tal manera que las hacemos desplazar entre un rango de valores manteniendo a la variable tiempo como fija. Por ejemplo para una línea en un tiempo $t = 30$ días de transcurrida su migración (ver Figura 5.8) se observa que la magnitud de la función de riesgo oscila entre 0.03 y 0.085.
En las figuras también se observa que a medida que el cargo fijo aumenta, el valor de la función de riesgo se incrementa mas rápido para los clientes jóvenes que para los clientes mayores. Estos resultados guardan coherencia con la realidad pues se espera que la mayoría de los clientes jóvenes tengan un poder adquisitivo más bajo que los clientes adultos, lo que no les permite adquirir un plan con un cargo fijo mayor.

Para apreciar de mejor manera el comportamiento anterior, graficamos los intervalos de confianza en el plano, fijando para ello a una de las covariables y generando un corte sobre la gráfica. Por ejemplo fijando a la variable edad, se observa que para una factura de S/.120 el valor de la función de riesgo subyacente se encuentra entre 0.04 y 0.055 para un individuo con una edad de 20 años mientras que para otro con 40 años el valor oscila entre 0.03 y 0.045. Para un valor de la factura de S/.150, la magnitud de la función se encuentra entre 0.05 y 0.085 y entre 0.04 y 0.075 respectivamente (ver Figura 5.9).

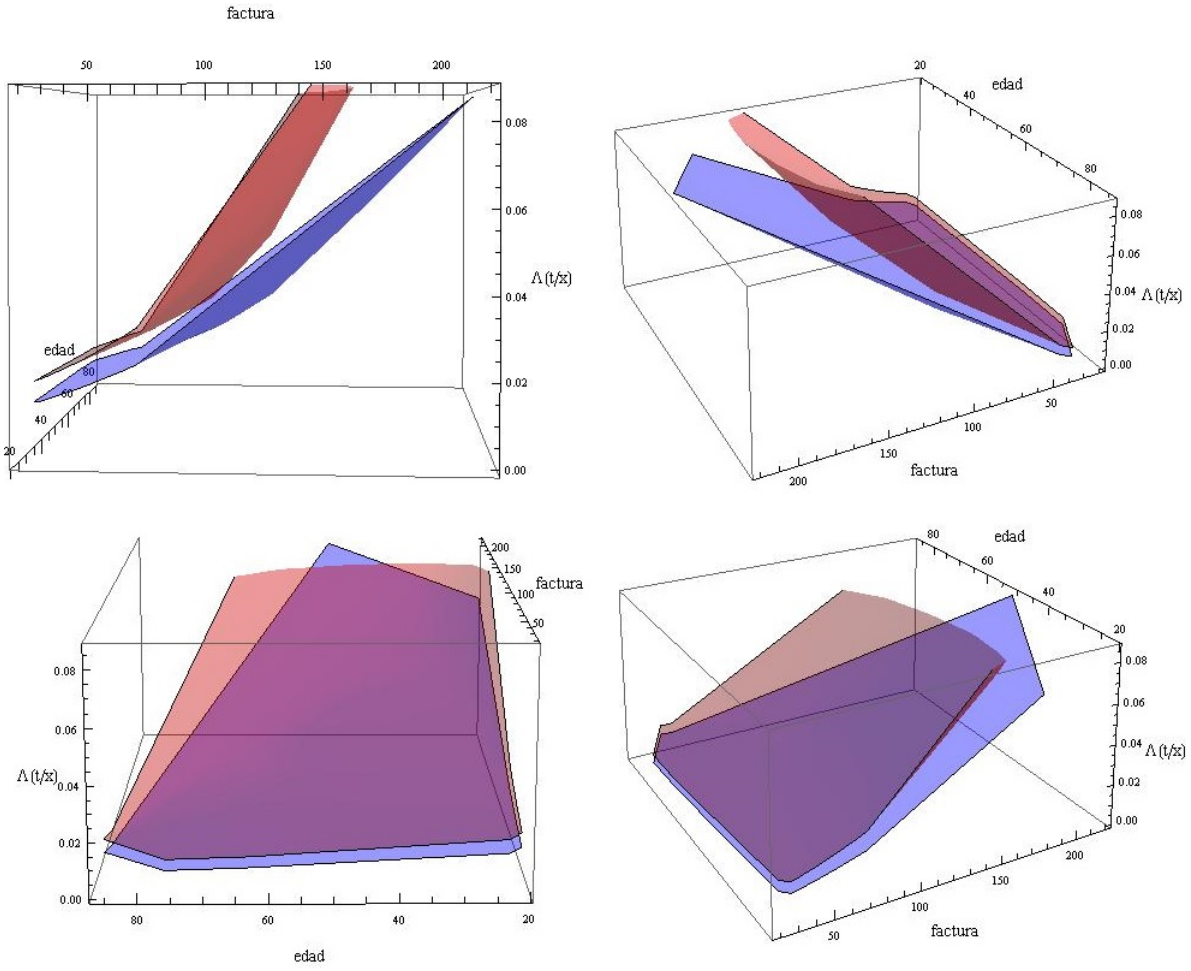

Figura 5.8: Intervalo de confianza para una línea transcurridos t=30 días de la migración (cuatro vistas distintas)

A continuación se muestran los escenarios para t = 90 y 180 días (figuras 5.10 y 5.11 respectivamente) en donde se llegan a observar similares características a las ya expuestas para t = 30 días.

El desarrollo de estos gráficos se encuentra en la sección (B.2) del apéndice B.

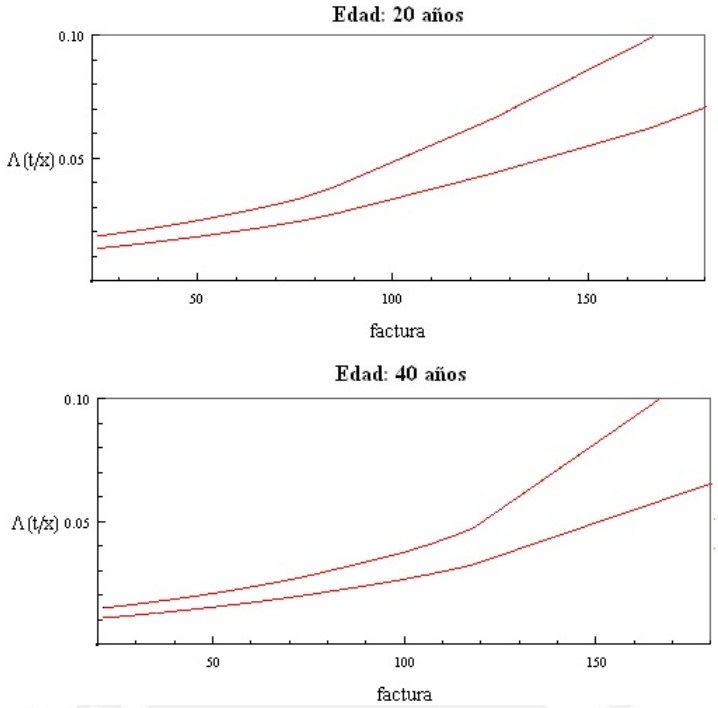

Figura 5.9: Riesgo asociado para un cliente con edad de 20 y 40 años respectivamente transcurridos t=30 días de la migración

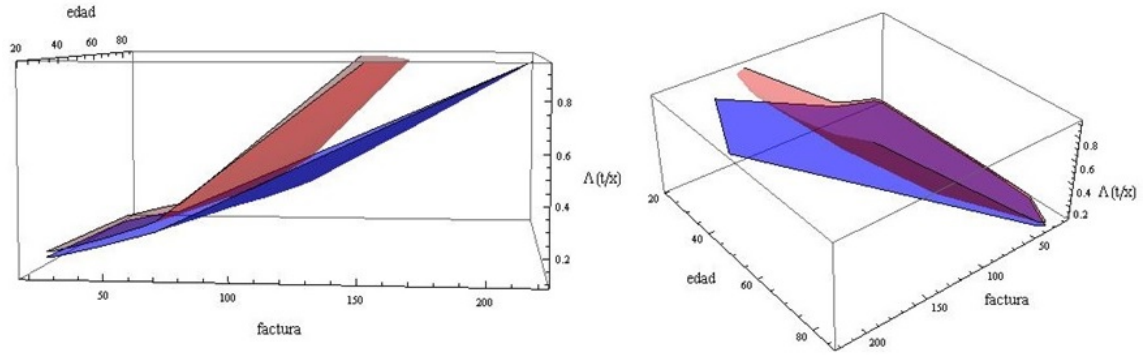

Figura 5.10: Intervalo de confianza para una línea transcurridos t=90 días de la migración

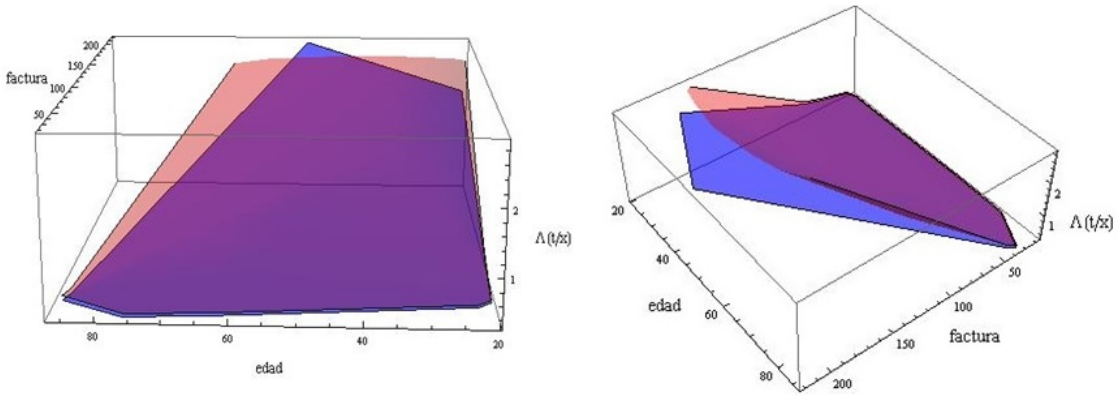

Figura 5.11: Intervalo de confianza para una línea transcurridos t=180 días de la migración

Capítulo 6

Conclusiones y sugerencias

6.1. Conclusiones

Del trabajo realizado se llegaron a las siguientes conclusiones:

- Cuando inicia el análisis, es decir en el mes de realizada la migración a Postpago, se observa que son pocas las líneas que se desactivan; sin embargo, conforme transcurre el tiempo, la cantidad de clientes que dan de baja a la línea se tiende a incrementar. Estos resultados guardan relación con la realidad. Se sabe que en los primeros meses de transcurrida la migración, pocas son las líneas que se desactivan (dan de baja a la línea o migran a Prepago). De acuerdo a estos resultados, al parecer existe una importante cantidad de clientes que no valoran el mantener su número telefónico (baja de la línea) o que no están dispuestos a mantener un pago fijo mensual por el servicio (migración a Prepago).
  Por otra parte, al considerar a las covariables factura y edad como libres y haciéndolas recorrer entre un rango de valores, se observa que a medida que transcurre el tiempo la magnitud de la función de azar se incrementa. También se puede concluir que a medida que el cargo fijo aumenta, el valor de la función de riesgo se incrementa mas rápido para los clientes jóvenes que para los clientes mayores, lo cual también tiene lógica con la realidad del mercado. A partir de ello se pueden ajustar los análisis de evaluación crediticia de los clientes de tal manera de que éstos influyan en su score.

- Un punto importante desarrollado respecto a los intervalos de confianza basados en pruebas estadísticas para cuantiles de supervivencia en el modelo de regresión de Cox es el uso de cuantiles bootstrap para aproximar los cuantiles de $\left\{\widehat{\Lambda}(t \mid x) - \Lambda(t \mid x)\right\} / \widehat{v}^{1/2}(x)$, en lugar de utilizar la aproximación normal (o ampliaciones de Edgeworth) como en trabajos anteriores de intervalos de confianza basados en pruebas estadísticas para ξ_p (en ausencia de covariables) a partir de datos de supervivencia con censura (Tze y Zheng (2006)).

- La novedad es que se trabaja con $\widehat{\Lambda}(t \mid x) - \log(1-p)^{-1}$, en lugar de $\widehat{S}(t \mid x) - (1-p)$ el cual ha sido utilizado por Brookmeyer y Crowley (1982) y autores posteriores para el caso en que no hay covariables. En presencia de covariables, hay una variabilidad adicional debido a la estimación del parámetro de regresión β, lo cual es útil para transformar $\widehat{S}(t \mid x)$, el cual esta restringido al intervalo $[0, 1]$, a diferencia del caso en

el que $\Lambda(t \mid x) - \log(1-p)^{-1}$ no esta restringido. Esta transformación, realizada por Tze y Zheng (2006), conduce a menudo a intervalos de confianza más cortos.

- Aunque $\widehat{S}(t \mid \boldsymbol{x})$ toma valores en el intervalo $[0,1]$, $\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{x}$ no tiene dicha limitación y su varianza en muestras finitas puede ser significativa. Por otra parte, la aproximación normal a $\left|\widehat{S}(t \mid \boldsymbol{x}) - S(t \mid \boldsymbol{x})\right| / \widehat{\sigma}(t \mid \boldsymbol{x})$ usada en la ecuación (3.1) puede ser inadecuada cuando el tamaño de la muestra no es lo suficientemente grande.

- En lugar de utilizar $\widehat{S}(t \mid \boldsymbol{x}) - (1-p)$ como la estadística de prueba, se usa la transformación logarítmica para convertirla en $\widehat{\Lambda}(t \mid \boldsymbol{x}) - \log(1-p)^{-1}$. Una ventaja de esta transformación es que a diferencia de $\widehat{S}(t \mid \boldsymbol{x})$, $\widehat{\Lambda}(t \mid \boldsymbol{x})$ ya no está limitada de pertenecer al intervalo $[0,1]$ y por lo tanto la variabilidad debida a $\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' x$ en su fórmula de varianza asintótica puede ser compatible con su magnitud.

- Bootstrap generalmente se aplica a funciones suavizadas de los datos y es muy útil para problemas donde la teoría asintótica de máxima verosimilitud no es fácil de aplicar.

- Un número grande de muestreos bootstrap es requerido cuando se tienen pocos individuos (tratamientos clínicos por ejemplo). Como se expuso en la segunda y tercera condición de regularidad de la sección (3.2) y dada la gran cantidad de datos con los que se cuenta, un número de $B = 100$ muestras bootstrap es sugerida para una buena precisión.
  Es importante mencionar también que las estimaciones de los parámetros mejoran al realizar los remuestreos.

6.2. Sugerencias para investigaciones futuras

- Por fines prácticos de tiempo, el estudio incluyó el uso de sólo dos covariables. Para un análisis mas riguroso respecto a la problemática de la desactivación de las lineas en la empresa objeto de estudio, se sugiere la adición de otras covariables como por ejemplo las características del plan tarifario es decir la cantidad de unidades de cada producto que brinda cada plan (X minutos a determinado destino, Y cantidad de megabytes, etc).
  El beneficio de agregar otras covariables se reflejará en conocer su relación con las dos ya consideradas y cómo en conjunto afectan en la desactivación de la linea. Con los resultados se podrá definir por ejemplo nuevos planes tarifarios con un determinado cargo fijo orientados a determinados segmentos de edad y con una determinada cantidad de minutos, mensajes de texto, datos, etc.

- En lugar de considerar a las dos covariables empleadas (edad y factura) como libres se pueden tomar otras como el tiempo, de tal manera que por ejemplo las dos primeras sean fijas o implícitas y la variable tiempo sea tomada como variable libre. Ello brindara otra perspectiva del problema.

- En lugar de emplear la estimación de Breslow se podría emplear otras estimaciones para la función de riesgo subyacente.

- Partiendo del análisis que obtuvimos, el trabajo podría extenderse al cálculo de bandas de confianza de la forma $\left\{\sqrt{n}\left(\widehat{\xi}_p(x) - \xi_p(x)\right), x \in K\right\}$ ya que converge débilmente a un proceso gaussiano indexado por $\boldsymbol{x} \in K$ cuando $n \to \infty$ donde K es un subconjunto compacto del espacio de covariables (Burr y Doss (1993)).

Apéndice A

Demostraciones y conceptos teóricos

A.1. Cálculo de la primera y segunda derivada de la función de máxima verosimilitud del modelo de Cox

A continuación se presenta el cálculo de la primera y segunda derivada de la función de máxima verosimilitud dada en la ecuación (1.2) que son empleadas para el cálculo de $\widehat{v}\left(t \mid x\right)$ en la ecuación (3.6). Así se tiene:

$$\begin{aligned} l\left(\beta\right) &= \sum_{i=1}^{n} \delta_i \left\{ \beta' x_i - \log\left(\sum_{j:\widetilde{t}_j \geq \widetilde{t}_i} \exp\left(\beta' x_j\right)\right)\right\} \\ &= \sum_{i=1}^{n} \delta_i \left\{ \left[\beta_0, \beta_1\right] \begin{bmatrix} x_{0i} \\ x_{1i} \end{bmatrix} - \log\left(\sum_{j:\widetilde{t}_j \geq \widetilde{t}_i} \exp\left(\left[\beta_0, \beta_1\right] \begin{bmatrix} x_{0i} \\ x_{1i} \end{bmatrix}\right)\right)\right\} \\ &= \sum_{i=1}^{n} \delta_i \left\{ \left[\beta_0 x_{0i} + \beta_1 x_{1i}\right] - \log\left(\sum_{j:\widetilde{t}_j \geq \widetilde{t}_i} \exp\left(\left[\beta_0 x_{0j} + \beta_1 x_{1j}\right]\right)\right)\right\} \end{aligned}$$

donde $\beta' = [\beta_0, \beta_1]$ y

$$x_i = \begin{bmatrix} x_{0i} \\ x_{1i} \end{bmatrix} = \begin{bmatrix} x_{01}, x_{02}, ..., x_{0n} \\ x_{11}, x_{12}, ..., x_{1n} \end{bmatrix}, \quad x_j = \begin{bmatrix} x_{0j} \\ x_{1j} \end{bmatrix} = \begin{bmatrix} x_{01}, x_{02}, ..., x_{0m} \\ x_{11}, x_{12}, ..., x_{1m} \end{bmatrix}.$$

Entonces se tiene que las derivadas parciales de primer orden respecto a cada parámetro son:

$$\begin{aligned} \frac{\partial l\left(\beta\right)}{\partial \beta_0} &= \sum_{i=1}^{n} \delta_i \left(x_{0i} - \frac{\left(\sum_{j:t_j > t_i} \exp\left(\left[\beta_0 x_{0j} + \beta_1 x_{1j}\right]\right)\right)'}{\sum_{j:t_j > t_i} \exp\left(\left[\beta_0 x_{0j} + \beta_1 x_{1j}\right]\right)}\right) \\ &= \sum_{i=1}^{n} \delta_i \left(x_{0i} - \frac{\sum_{j:t_j > t_i} x_{0j} \exp\left(\left[\beta_0 x_{0j} + \beta_1 x_{1j}\right]\right)}{\sum_{j:t_j > t_i} \exp\left(\left[\beta_0 x_{0j} + \beta_1 x_{1j}\right]\right)}\right) \end{aligned}$$

$$\frac{\partial l\left(\beta\right)}{\partial \beta_1} = \sum_{i=1}^{n} \delta_i \left(x_{1i} - \frac{\sum_{j:t_j > t_i} x_{1j} \exp\left(\left[\beta_0 x_{0j} + \beta_1 x_{1j}\right]\right)}{\sum_{j:t_j > t_i} \exp\left(\left[\beta_0 x_{1j} + \beta_1 x_{1j}\right]\right)}\right)$$

Luego, las derivadas parciales de segundo orden respecto a cada parámetro son:

$$\frac{\partial l^2(\beta)}{\partial {\beta_0}^2} = -\sum_{i=1}^{n} \delta_i \left\{ \frac{\left[\sum_{j:t_j>t_i} x_{0j}^2 \exp\left([\beta_0,\beta_1]\, x_j\right)\right]\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]}{\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2} - \frac{\left[\sum_{j:t_j>t_i} x_{0j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]\left[\sum_{j:t_j>t_i} x_{0j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]}{\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2} \right\}$$

$$= -\sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{j:t_j>t_i} x_{0j}^2 \exp\left([\beta_0,\beta_1]\, x_j\right)}{\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)} - \frac{\left[\sum_{j:t_j>t_i} x_{0j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2}{\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2} \right\}$$

$$\frac{\partial l^2(\beta)}{\partial {\beta_1}^2} = -\sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{j:t_j>t_i} x_{1j}^2 \exp\left([\beta_0,\beta_1]\, x_j\right)}{\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)} - \frac{\left[\sum_{j:t_j>t_i} x_{1j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2}{\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2} \right\}$$

$$\frac{\partial l^2(\beta)}{\partial \beta_0 \partial \beta_1} = -\sum_{i=1}^{n} \delta_i \left\{ \frac{\left[\sum_{j:t_j>t_i} x_{0j}x_{1j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]}{\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2} - \frac{\left[\sum_{j:t_j>t_i} x_{1j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]\left[\sum_{j:t_j>t_i} x_{0j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]}{\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2} \right\}$$

$$= -\sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{j:t_j>t_i} x_{0j}x_{1j} \exp\left([\beta_0,\beta_1]\, x_j\right)}{\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)} - \frac{\left[\sum_{j:t_j>t_i} x_{1j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]\left[\sum_{j:t_j>t_i} x_{0j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]}{\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2} \right\}$$

$$\frac{\partial l^2(\beta)}{\partial \beta_1 \partial \beta_0} = -\sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{j:t_j>t_i} x_{1j}x_{0j} \exp\left([\beta_0,\beta_1]\, x_j\right)}{\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)} - \frac{\left[\sum_{j:t_j>t_i} x_{0j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]\left[\sum_{j:t_j>t_i} x_{1j} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]}{\left[\sum_{j:t_j>t_i} \exp\left([\beta_0,\beta_1]\, x_j\right)\right]^2} \right\}$$

Como complemento se expone un concepto citado durante el desarrollo del trabajo:

A.2. Expansiones Edgeworth

Expuesta por Frances Ysidro Edgeworth, quien conociendo que muchos datos reales no están normalmente distribuidos, introdujo una expansión de series, la cual, rivalizando con la familia de Pearson, proporciona flexibilidad en la descripción de la simetría y otros fenómenos. Esto es lo que se conoce como expansión de Edgeworth, que desde entonces se ha convertido en un instrumento útil tanto para la Econometría como para la Estadística, mejorando las aproximaciones de las distribuciones en el muestreo que ofrece el Teorema del Límite Central. Son series que aproximan una distribución de probabilidad en términos de sus cumulantes. Los cumulantes de una distribución de probabilidad son un conjunto de cantidades que proporcionan una alternativa a los momentos de la distribución. Los momentos determinan los cumulantes en el sentido de que cualquiera de dos distribuciones de probabilidad cuyos momentos sean idénticos tendrán también cumulantes idénticos, y de manera similar los cumulantes determinan los momentos.
Edgeworth (1905) desarrolló una expansión similar a partir del Teorema del Límite Central. La ventaja de la serie es que se controla el error, siendo así una verdadera serie asintótica.

Weisstein (2013) propone la siguiente definición: Sea una distribución que se aproxima a la distribución F_n de sumas estandarizadas

$$Y_n = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)}{\sqrt{\sum_{i=1}^{n} \sigma_x^2}}. \tag{A.1}$$

En las series Charlier, se toman las variables aleatorias i.i.d. con media μ, varianza σ^2 y cumulantes mayores $\sigma^r \lambda_r$ para $r \geq 3$. También, se toma a $\Psi(t)$ como la función de distribución normal estándar $\Phi(t)$, por lo que tenemos:

$$\begin{aligned} k_1 - \gamma_1 &= 0, \\ k_2 - \gamma_2 &= 0, \\ k_3 - \gamma_3 &= \frac{\lambda^r}{n^{r/2} - 1}. \end{aligned}$$

Entonces la serie Edgeworth es obtenida por agrupación de términos para obtener la expansión asintótica de la función característica de la forma:

$$f_n(t) = \left[1 + \sum_{r=1}^{\infty} \frac{P_r(it)}{n^{r/2}}\right] \exp^{-r^2/2}, \tag{A.2}$$

donde P_r es un polinomio de grado 3r con los coeficientes en función de los cumulantes de orden de 3 a $r + 2$. Si las potencias de Ψ se consideran como las derivadas, entonces la extensión de la función de distribución está dada por:

$$F_n(x) = \Psi(x) + \sum_{r=1}^{\infty} \frac{P_r(-\Phi(x))}{n^{r/2}} \tag{A.3}$$

(Wallace (1958)). Los primeros términos de esta expansión son dadas por:

$$f(t) = \Psi(t) - \frac{\lambda_3 \Psi^{(3)}(t)}{6\sqrt{n}} + \frac{1}{n}\left[\frac{\lambda_4 \Psi^{(4)}(t)}{24} + \frac{\lambda_3^2 \Psi^{(6)}(t)}{72}\right] + ... \tag{A.4}$$

Cramér (1928) prueba que ésta serie es uniformemente válida en t.

Apéndice B

Rutinas: Códigos de programas

B.1. Cálculo de los intervalos de confianza

A continuación se muestra el código asociado para el cálculo de los parámetros del modelo de Cox en el software *R*. Como fue mencionado en el punto 5.2, también fue desarrollado en los software *Mathematica* y *SPSS*. Luego se muestra el código que corresponde al cálculo de los parámetros para las B muestras bootstrap. Ambos pasos son necesarios para el cálculo de los intervalos de confianza.

```
#########################################################################
##Cálculo de los parámetros del modelo de Cox para la muestra original

#carga de librerias
library(survival)
require(survival)
library(psych)
require(psych)

#lectura de datos
me=read.table("C://Users//Jorge//Documents//ejm mathematica//ejm_ME_15K.txt", header=T)
me
#carga en memoria
attach(me)
#tipo de objeto creado
objects(2)
class(me)

#histograma de las dos covariables del modelo
hist(EDAD, main="Histograma de la variable Edad", xlab="variable Edad", ylab="Frecuencia",
breaks=19, col="green", border="black")
hist(FACTURA, main="Histograma de la variable Factura", xlab="variable Factura",
ylab="Frecuencia", breaks=19, col="green", border="black")

#cálculo del modelo de Cox
cox1<-coxph(Surv(TIEMPO,ESTADO)~FACTURA+EDAD, data=me, na.action=na.exclude, init=c(1,1))
summary(cox1)

#########################################################################
##Cálculo de los parámetros del modelo de Cox para las B muestras bootstrap

#se define la cantidad de muestras bootstrap (B) y la dimensión de la matriz que almacena
los parámetros de cada muestreo.
B=100;
mat <- matrix(,B,2)
```

```
#se construye la ruta donde se leen los archivos que contienen los "n" datos de c/u de
las B muestras bootstrap. Las muestras son generadas en el software Mathematica.
a="a";a0="C://Users//Jorge//Documents//files tesis//";a1="dB";a2=".csv"
for(i in 1:B) { a[i]<-paste(a0,a1,i,a2,sep="") }
a[1];a[2];a[3];

#se construye la entrada de los dos parámetros a estimar para cada muestreo bootstrap
b="b";C=2;b0="cox1\$coef[";b1="]"
for(i in 1:C) { b[i]<-paste(b0,i,b1,sep="") }
b[1];b[2];

#proceso iterativo que almacena en un archivo los parámetros estimados para cada muestreo
bootstrap
for(i in 1:B) {
a[i]<-paste(a0,a1,i,a2,sep="");
me=read.csv(a[i], fill = TRUE, header = FALSE);
ID=me\$V1;TIEMPO=me\$V2;ESTADO=me\$V3;FACTURA=me\$V4;EDAD=me\$V5;
cox1<-coxph(Surv(TIEMPO,ESTADO)~FACTURA+EDAD, data=me, na.action=na.exclude, init=c(1,1));
mat[i,1]=paste(cox1\$coef[1]);
mat[i,2]=paste(cox1\$coef[2]);
}

#se almacena en un archivo los parámetros de c/u de las muestras bootstrap para que sea leído
desde el software Mathematica
write.table(mat, file = "betas.csv", sep = ",", row.names = FALSE, col.names = FALSE)
```

Los códigos asociados para el desarrollo de los diagrama de flujo para el cálculo de los intervalos de confianza y el cálculo de los cuantiles bootstrap fueron diseñados en el software *Mathematica*. A continuación se exponen dichos códigos dados en el capítulo 4 (Figuras 4.1, 4.2, 4.3 y 4.4).

```
(*INTERVALOS DE CONFIANZA PARA LA MEDIANA DE SUPERVIVENCIA EN EL MODELO DE REGRESION DE COX*)
(*********************************************************************************************)
(*1. Generación de Muestras y asignación de variables *)

(*Limpieza de variables*)
Clear[n, datos, B, dB, datosB, IDB, TIEMPOB, ESTADOB, x1B, x2B, e, \[Beta]\[Beta], i, j, k,
xx, txx, c, t, q, tii, ti, \[Delta]];

(*Lectura de datos*)
datos := Import["C:/Users/Jorge/Documents/ejm mathematica/ejm_ME_15K.txt","Table"]

(*Tamaño de los datos*)
n = Length[datos];

(*Definición del número de muestras a generar/procesar*)
Clear[dB, datosB]; B = 100;

(*Generación de muestras bootstrap*)
(*Se crean B archivos con las muestras generadas y los almacena en un directorio para que
luego desde el R se calculen los parámetros asociados a cada muestra bootstrap*)
dB := Array[datosB, B];
For[k = 1, k <= B, k++, datosB[k] = RandomChoice[datos, Length[datos]]];
For[k = 1, k <= B, k++, IDB[k] = dB[[k]][[All, 1]]; TIEMPOB[k] = dB[[k]][[All, 2]];
 ESTADOB[k] = dB[[k]][[All, 3]];  x1B[k] = dB[[k]][[All, 4]]; x2B[k] = dB[[k]][[All, 5]]];

(*Exportación de las muestras generadas*)
For[k = 1, k <= B, k++, e = {"dB"}~Join~{k}~Join~{".csv"};
```

```
Export[ToString[e[[1]]] <> ToString[e[[2]]] <> ToString[e[[3]]], datosB[k]]];

(*Asignación de las variables. Se definen vectores covariables x1=FACTURA, x2=EDAD)*)
dB := Array[datosB, B];
For[k = 1, k <= B, k++, f = {"C:/Users/Jorge/Documents/dB"}~Join~{k}~Join~{".csv"};
  datosB[k] = Import[ToString[f[[1]]] <> ToString[f[[2]]] <> ToString[f[[3]]]]];
For[k = 1, k <= B, k++, IDB[k] = dB[[k]][[All, 1]]; TIEMPOB[k] = dB[[k]][[All, 2]];
ESTADOB[k] = dB[[k]][[All, 3]]; x1B[k] = dB[[k]][[All, 4]]; x2B[k] = dB[[k]][[All, 5]]];

(*Se importan los betas calculados para c/u de las B muestras mediante el Modelo de Cox*)
\[Beta]\[Beta] := Import["betas1.csv"];

(* Se agrupa el vector de covariables: x1,x2*)
For[k = 1, k <= B, k++, xx[k] = {x1B[k], x2B[k]}; txx[k] = Transpose[xx[k]]];

(*cálculo de la terna (ti,\[Delta]i,xi)*)
(*Definición de ci*)
c = 365;
(*cálculo de ti*)
For[k = 1, k <= B, k++, t[k] = TIEMPOB[k]]; q[i_, k_] := Min[t[k][[i]], c];
 tii = Table[q[i, k], {i, n}, {k, B}];
(*Asignación de la variable ESTADO a \[Delta]i*)
For[k = 1, k <= B, k++, \[Delta][k] = ESTADOB[k]];
(*Formación de la terna*)
Clear[terna, i]; terna[i_] := {ti[[i]], \[Delta][[i]], xx[[All, i]]};
ternat := For[i = 1, i <= n, i++, Print[terna[i]]];

(*****************************************************************************************)
(* 2. Cálculo de la Función Hazard Base Acumulada y características asociadas *)

(*Limpieza de variables*)
Clear[m, datos1, ID0, TIEMPO0, ESTADO0, x10, x20, \[Beta]0, xx0, x10, x20, txx0, xx0, t0,
\[Delta]0, ci0, p0, q0, ti0, tt0, pp00, CCC, CCCI, ccc, WWejm, WW, pp140, \[Delta]0,
\[CapitalLambda]t0, pp10, \[CapitalLambda]txx0, Lambda0, MStxx0, ppp30, W10, Wt10, W20,
Wt20, a, x, y, Q10, Qta10, QQ10, Q20, Qta20, QQ20, Qtxx0, pp100, pp200, PD0, pp150, pp0150,
pp1150, pp160, pp1160, SD110, SD220, SD210, SD0, ISD0, pp170, pp1170, pp11170, pp190,
pp1190, pp11190, pp2000, acum40, pp270, vtxx0, Mvt0];

(*Asignación de las variables del archivo (vectores covariable x1=FACTURA, x2=EDAD)*)
datos1 = Transpose[datos]; ID0 = datos1[[1]]; TIEMPO0 = datos1[[2]]; ESTADO0 = datos1[[3]];
x10 = datos1[[4]]; x20 = datos1[[5]];
\[Beta]0 = {0.011627, -0.006216};
xx0 = {x10, x20}; txx0 = Transpose[xx0]; t0 = TIEMPO0; \[Delta]0 = ESTADO0;
(*cálculo del vector ci*)
c = 365; ci0 = Array[p0, n]; For[i = 1, i <= n, i++, p0[i] = c];
(*cálculo de ti*)
q0[i_] := Min[t0[[i]], 365]; ti0 = Table[q0[i], {i, n}];
(*Definición del vector de tiempo a evaluar (discreto): 12meses de 30 dias c/u*)
m = 12; tt0 = Array[pp00, m]; For[i = 1, i <= m - 1, i++, pp00[i] = i*30; pp00[12] = c];

(*cálculo de los indices*)
Clear[CCC, CCCI, ccc]; CCCI = Array[ccc, n]; For[i = 1, i <= n, i++, ccc[i] = {}];
(*función que genera los indices*)
CCC[i_] := For[j = 1, j <= n, j++, If[ti0[[j]] >= ti0[[i]], ccc[i] = Union[ccc[i], {j}]]];
(*ejecucion de la funcion que genera los indices*)
Do[CCC[i], {i, 1, n}]; Export["CCCI.txt", CCCI, "Table"];

(*cálculo de W(t)*)
Clear[WWejm];
WWejm[j_] := Sum[Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}];
WW = Table[WWejm[j], {j, 1, n}];
Export["WW.txt", WW, "Table"];

(*cálculo de la función de azar base acumulada \[CapitalLambda](t)*)
```

```
For[j = 1, j <= m, j++, pp140[j] = 0]; \[CapitalLambda]t0 := Array[pp140, m];
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++,
   If[ti0[[i]] <= tt0[[j]], pp140[j] = \[Delta]0[[i]]/WW[[i]] + pp140[j]]]];
MatrixForm[N[\[CapitalLambda]t0, 2]]
Export["\[CapitalLambda]t0.txt", \[CapitalLambda]t0, "Table"];

(*cálculo de la función de azar acumulada \[CapitalLambda](t/x)*)
Clear[pp10, \[CapitalLambda]txx0];
pp10[j_, i_] := \[CapitalLambda]t0[[j]] Exp[\[Beta]0.txx0[[i, All]]];
\[CapitalLambda]txx0 = Table[pp10[j, i], {j, 1, m}, {i, 1, n}];
MatrixForm[ N[\[CapitalLambda]txx0, 2]] (*\[CapitalLambda]txx0=Lambda0*)
Lambda0 = \[CapitalLambda]txx0; MatrixForm[Lambda0];
Export["Lambda0.txt", Lambda0, "Table"];

(*cálculo de la función de supervivencia acumulada S(t/x)*)
Clear[MStxx0]; For[i = 1, i <= n, i++,
 For[j = 1, j <= m, j++, ppp30[j, i] = 0]]; MStxx0 := Array[ppp30, {m, n}];
For[i = 1, i <= n, i++, For[j = 1, j <= m, j++, ppp30[j, i] = Exp[-pp10[j, i]]]];
MatrixForm[N[MStxx0, 2]]
Export["MStxx0.txt", MStxx0, "Table"];

(*cálculo de Wl(t)*)
(*para l=1*)
Clear[W10, Wt10];
Wt10[j_] := Sum[txx0[[i]][[1]]*Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}];
W10 = Table[Wt10[j], {j, 1, n}];
MatrixForm[N[W10, 2]];
(*para l=2*)
Clear[W20, Wt20];
Wt20[j_] := Sum[txx0[[i]][[2]]*Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}];
W20 = Table[Wt20[j], {j, 1, n}];
MatrixForm[N[W20, 2]];

(*Definición de a ("a" es vector)*)
Clear[a, x, y]; a = {x[1], y[2]};

(*cálculo de Ql(t,a)*)
(*para l=1*)
Clear[Q10, Qta10, QQ10]; Qta10 := Array[Q10, m]; For[j = 1, j <= m, j++, Q10[j] = 0];
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++,
   If[ti0[[i]] <= tt0[[j]], Q10[j] = \[Delta]0[[i]]*(Wt10[i]/WW[[i]] - x)/WW[[i]] + Q10[j]]]];
MatrixForm[N[Qta10, 2]]
Export["Qta10.txt", Qta10, "Table"];
(*para l=2*)
Clear[Q20, Qta20, QQ20]; Qta20 := Array[Q20, m]; For[j = 1, j <= m, j++, Q20[j] = 0];
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++,
   If[ti0[[i]] <= tt0[[j]], Q20[j] = \[Delta]0[[i]]*(Wt20[i]/WW[[i]] - y)/WW[[i]] + Q20[j]]]];
MatrixForm[N[Qta20, 2]]
Export["Qta20.txt", Qta20, "Table"];

(*Formación de Q(t,x)*)
Clear[Qtxx0]; Qtxx0[x_, y_] = {QQ10[x], QQ20[y]};
MatrixForm[N[Qtxx0[x, y]]]; MatrixForm[Transpose[Qtxx0[x, y]]];
Dimensions[Qtxx0[x, y]]; Dimensions[Transpose[Qtxx0[x, y]]];

(*cálculo de l..(b): primera y segunda derivada*)
(*cálculo de la 1ra derivada*)
Clear[pp100, pp200, PD0];
For[j = 1, j <= n, j++, pp150[j] = 0]; For[j = 1, j <= n, j++, pp0150[j] = 0];
For[j = 1, j <= n, j++, pp1150[j] = 0]; pp160 = 0; pp1160 = 0; pp100 = 0; pp200 = 0;
(*para beta0*)
For[j = 1, j <= n, j++, pp150[j] = Sum[txx0[[i]][[1]] Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}]];
For[j = 1, j <= n, j++, pp0150[j] = Sum[Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}]];
pp160 = Sum[\[Delta]0[[i]] txx0[[i]][[1]] - \[Delta]0[[i]] pp150[i]/pp0150[i], {i, n}];
```

```
pp100 = pp160;
(*para beta1*)
For[j = 1, j <= n, j++, pp1150[j] = Sum[txx0[[i]][[2]] Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}]];
pp1160 = Sum[\[Delta]0[[i]] txx0[[i]][[2]] - \[Delta]0[[i]] pp1150[i]/pp0150[i], {i, n}];
pp200 = pp1160;
(*primera derivada*)
PD0 = {pp100, pp200};
MatrixForm[N[PD0, 4]]

(*cálculo de la 2da derivada*)
Clear[SD110, SD220, SD210, SD0, pp170, pp1170, pp11170, pp190, pp1190, pp11190];
For[j = 1, j <= m, j++, pp170[j] = 0]; For[j = 1, j <= m, j++, pp190[j] = 0]; SD110 = 0;
For[j = 1, j <= m, j++, pp1170[j] = 0]; For[j = 1, j <= m, j++, pp1190[j] = 0]; SD220 = 0;
For[j = 1, j <= m, j++, pp11170[j] = 0]; For[j = 1, j <= m, j++, pp11190[j] = 0]; SD210 = 0;
(*elemento 11*)
For[j = 1, j <= n, j++,
pp170[j] = Sum[(txx0[[i]][[1]]^2) Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}]];
pp190 = Sum[-\[Delta]0[[i]] (pp170[i]/pp0150[i] - ((pp150[i]/pp0150[i])^2)), {i, n}];
SD110 := pp190;
(*elemento 22*)
For[j = 1, j <= n, j++,
pp1170[j] = Sum[(txx0[[i]][[2]]^2) Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}]];
pp1190 = Sum[-\[Delta]0[[i]] (pp1170[i]/pp0150[i] - ((pp1150[i]/pp0150[i])^2)), {i, n}];
SD220 := pp1190;
(*elemento 12=21*)
For[j = 1, j <= n, j++,
  pp11170[j] = Sum[(txx0[[i]][[1]] txx0[[i]][[2]]) Exp[\[Beta]0.txx0[[i]]], {i, ccc[j]}]];
pp11190 =
Sum[-\[Delta]0[[i]] (pp11170[i]/pp0150[i] - pp150[i] pp1150[i]/(pp0150[i]^2)), {i, n}];
SD210 := pp11190;
(*Forma de la matriz de la 2da derivada*)
SD0 = {{SD110, SD210}, {SD210, SD220}};
MatrixForm[N[SD0, 4]]
Export["SD0.txt", SD0, "Table"];

(*Inversa y cálculo del producto*)
Clear[ISD0]; ISD0 = Inverse[(-1)*SD0];
MatrixForm[N[ISD0, 2]]
MatrixForm[SD0.ISD0]

(*cálculo de v(t/x)*)
(*primer miembro del componente v(t/x)*)
For[j = 1, j <= m, j++, pp2000[j] = 0]; acum40 := Array[pp2000, m];
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++,
   If[ti0[[i]] <= tt0[[j]], pp2000[j] = \[Delta]0[[i]]/(WW[[i]]^2) + pp2000[j]]]];
MatrixForm[N[acum40, 2]]
Export["acum40.txt", acum40, "Table"];

(*segundo miembro del componente v(t/x)*)
Clear[pp270];
pp270[j_,i_] :=
(Qtxx0[xx0[[1, i]], xx0[[2, i]]][[All, j]].ISD0.Qtxx0[xx0[[1, i]], xx0[[2, i]]][[All, j]]);

(*cálculo de v(t/x)*)
Clear[Mvt0]; For[j = 1, j <= m, j++, For[i = 1, i <= n, i++, vtxx0[j, i] = 0] ];
For[j = 1, j <= m, j++, Mvt0 := Array[vtxx0, {m, n}];
 For[i = 1, i <= n, i++, vtxx0[j, i] = Exp[2*\[Beta]0.txx0[[i]]]*(pp2000[j] + pp270[j, i])]];
MatrixForm[Mvt0]; Export["Mvt0.txt", Mvt0, "Table"];

(*cálculo de \[Delta]2(t/x)*)
MStxxx0=MStxx0; For[i = 1, i <= n, i++, For[j = 1, j <= m, j++, pp30[j, i] = 0]];
\[Delta]2tx := Array[pp30, {m, n}];
For[i = 1, i <= n, i++, For[j = 1, j <= m, j++,
pp30[j, i] = (N[MStxxx0[[j]][[i]], 2]^2)*Mvtt0[[j]][[i]]]];
```

```
MatrixForm[N[\[Delta]2tx, 2]]

(*****************************************************************************************)
(* 3. Cálculo de las características de la b-ésima muestra bootstrap *)

(*Limpieza de variables*)
Clear[b, ID, TIEMPO, ESTADO, x1, x2, \[Beta], tt, pp0, ti, CC, CCI, cc, Wejm, W,
\[CapitalLambda]t, pp14, ppp1, MAtxx, MStxx, ppp3, W1, Wt1, W2, Wt2, a, x, y, Q1, Qta1, QQ1,
Q2, Qta2, QQ2, Qtxx, pp1, pp2, PD, pp15, pp015, pp115, pp16, pp116, PD, SD11, SD22, SD21, SD,
pp17, pp117, pp1117, pp19, pp119, pp1119, ISD, pp20, acum4, pp27, vtxx, Mvt];

(*Lectura de datos*)
(*se define el nro de muestra con reemplazo (B)*)
b = 100;
ID = datosB[b][[All, 1]]; TIEMPO = datosB[b][[All, 2]]; ESTADO = datosB[b][[All, 3]];
\[Delta] = ESTADO; x1 = datosB[b][[All, 4]]; x2 = datosB[b][[All, 5]]; ti = tii[[All, b]];
xx = {x1, x2}; txx = Transpose[xx];
(*Se leen los parametros estimados mediante el Modelo de Cox para cada muestra bootstrap*)
\[Beta] = \[Beta]\[Beta][[b, All]]

(*Definición del vector tt (tiempo por evaluar: 12 de 30 días c/u)*)
m = 12; tt := Array[pp0, m];
For[j = 1, j <= m - 1, j++, pp0[j] = j*30; pp0[12] = c];

(*cálculo de los indices*)
Clear[CC, CCI, cc]; CCI = Array[cc, n]; For[i = 1, i <= n, i++, cc[i] = {}];
(*funcion que genera los indices*)
CC[i_] := For[j = 1, j <= n, j++, If[ti[[j]] >= ti[[i]], cc[i] = Union[cc[i], {j}]]];
(*ejecución de la función que genera los indices*)
Do[CC[i], {i, 1, n}]; Timing[Do[CC[i],{i,1,n}];

(*cálculo de W(t)*)
Clear[Wejm, W];
Wejm[j_] := Sum[Exp[\[Beta].txx[[i]]], {i, cc[j]}];
W = Table[Wejm[j], {j, 1, n}];

(*cálculo de la función de azar base acumulada \[CapitalLambda](t)*)
For[j = 1, j <= m, j++, pp14[j] = 0];
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++,
   If[ti[[i]] <= tt[[j]], pp14[j] = \[Delta][[i]]/W[[i]] + pp14[j]]]];
\[CapitalLambda]t = Array[pp14, m]; MatrixForm[\[CapitalLambda]t]

(*cálculo de la función de azar acumulada \[CapitalLambda](t/x)*)
Clear[ppp1, MAtxx];
ppp1[j_, i_] := \[CapitalLambda]t[[j]] Exp[\[Beta].txx[[i, All]]];
MAtxx = Table[ppp1[j, i], {j, 1, m}, {i, 1, n}];
MatrixForm[N[MAtxx, 2]]

(*cálculo de la función de supervivencia acumulada (Stx)*)
Clear[MStxx]; For[i = 1, i <= n, i++, For[j = 1, j <= m, j++, ppp3[j, i] = 0]];
MStxx := Array[ppp3, {m, n}];
For[i = 1, i <= n, i++, For[j = 1, j <= m, j++, ppp3[j, i] = Exp[-ppp1[j, i]]]];
MatrixForm[N[MStxx, 2]]

(*cálculo de Wl(t)*)
(*para l=1*)
Clear[W1, Wt1];
Wt1[j_] := Sum[txx[[i]][[1]]*Exp[\[Beta].txx[[i]]], {i, cc[j]}];
W1 = Table[Wt1[j], {j, 1, n}];
(*para l=2*)
Clear[W2, Wt2];
Wt2[j_] := Sum[txx[[i]][[2]]*Exp[\[Beta].txx[[i]]], {i, cc[j]}];
W2 = Table[Wt2[j], {j, 1, n}];
```

```
(*Definición de a ("a" es vector)*)
Clear[a, x, y]; a = {x[1], y[2]};

(*cálculo de Ql(t,a)*)
(*para l=1*)
Clear[Q1, Qta1, QQ1]; Qta1 := Array[Q1, m]; For[j = 1, j <= m, j++, Q1[j] = 0];
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++,
   If[ti[[i]] <= tt[[j]], Q1[j] = \[Delta][[i]]*(Wt1[i]/W[[i]] - x)/W[[i]] + Q1[j]]]];
MatrixForm[N[Qta1, 2]];
(*para l=2*)
Clear[Q2, Qta2, QQ2]; Qta2 := Array[Q2, m]; For[j = 1, j <= m, j++, Q2[j] = 0];
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++,
   If[ti[[i]] <= tt[[j]], Q2[j] = \[Delta][[i]]*(Wt2[i]/W[[i]] - y)/W[[i]] + Q2[j]]]];
MatrixForm[N[Qta2, 2]];

(*Formación de Q(t,x)*)
Clear[Qtxx]; Qtxx[x_, y_] = {QQ1[x], QQ2[y]};
MatrixForm[N[Qtxx[x, y]]]; MatrixForm[Transpose[Qtxx[x, y]]];
Dimensions[Qtxx[x, y]]; Dimensions[Transpose[Qtxx[x, y]]];

(*cálculo de l..(b): primera y segunda derivada*)
Clear[pp1, pp2, PD];
(*cálculo de la 1ra derivada*)
For[j = 1, j <= n, j++, pp15[j] = 0]; For[j = 1, j <= n, j++, pp015[j] = 0];
For[j = 1, j <= n, j++, pp115[j] = 0]; pp16 = 0; pp116 = 0; pp1 = 0; pp2 = 0;
(*para beta0*)
For[j = 1, j <= n, j++, pp15[j] = Sum[txx[[i]][[1]] Exp[\[Beta].txx[[i]]], {i, cc[j]}]];
For[j = 1, j <= n, j++, pp015[j] = Sum[Exp[\[Beta].txx[[i]]], {i, cc[j]}]];
pp16 = Sum[\[Delta][[i]] txx[[i]][[1]] - \[Delta][[i]] pp15[i]/pp015[i], {i, n}];
pp1 = pp16;
(*para beta1*)
For[j = 1, j <= n, j++, pp115[j] = Sum[txx[[i]][[2]] Exp[\[Beta].txx[[i]]], {i, cc[j]}]];
pp116 = Sum[\[Delta][[i]] txx[[i]][[2]] - \[Delta][[i]] pp115[i]/pp015[i], {i, n}];
pp2 = pp116;
(*primera derivada*)
PD = {pp1, pp2};
MatrixForm[N[PD, 4]]

(*cálculo de la 2da derivada*)
Clear[SD11, SD22, SD21, SD, pp17, pp117, pp1117, pp19, pp119, pp1119];
For[j = 1, j <= m, j++, pp17[j] = 0]; For[j = 1, j <= m, j++, pp19[j] = 0]; SD11 = 0;
For[j = 1, j <= m, j++, pp117[j] = 0]; For[j = 1, j <= m, j++, pp119[j] = 0]; SD22 = 0;
For[j = 1, j <= m, j++, pp1117[j] = 0]; For[j = 1, j <= m, j++, pp1119[j] = 0]; SD21 = 0;
(*elemento 11*)
For[j = 1, j <= n, j++,
  pp17[j] = Sum[(txx[[i]][[1]]^2) Exp[\[Beta].txx[[i]]], {i, cc[j]}]];
pp19 = Sum[-\[Delta][[i]] (pp17[i]/pp015[i] - ((pp15[i]/pp015[i])^2)), {i, n}];
SD11 := pp19;
(*elemento 22*)
For[j = 1, j <= n, j++,
  pp117[j] = Sum[(txx[[i]][[2]]^2) Exp[\[Beta].txx[[i]]], {i, cc[j]}]];
pp119 = Sum[-\[Delta][[i]] (pp117[i]/pp015[i] - ((pp115[i]/pp015[i])^2)), {i, n}];
SD22 := pp119;
(*elemento 12=21*)
For[j = 1, j <= n, j++,
  pp1117[j] = Sum[(txx[[i]][[1]] txx[[i]][[2]]) Exp[\[Beta].txx[[i]]], {i, cc[j]}]];
pp1119 = Sum[-\[Delta][[i]] (pp1117[i]/pp015[i] - pp15[i] pp115[i]/(pp015[i]^2)), {i, n}];
SD21 := pp1119;
(*Forma de la matriz de la 2da derivada*)
SD = {{SD11, SD21}, {SD21, SD22}};
MatrixForm[N[SD, 4]]

(*Inversa y cálculo del producto*)
Clear[ISD];ISD = Inverse[(-1)*SD];
```

```
MatrixForm[N[ISD, 2]]
MatrixForm[SD.ISD]

(*cálculo de v(t/x)*)
(*primer miembro del componente v(t/x)*)
For[j = 1, j <= m, j++, pp20[j] = 0]; acum4 := Array[pp20, m];
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++,
   If[ti[[i]] <= tt[[j]], pp20[j] = \[Delta][[i]]/(W[[i]]^2) + pp20[j]]]];
MatrixForm[N[acum4, 2]]

(*segundo miembro del componente v(t/x)*)
Clear[pp27];
pp27[j_, i_] :=
(Qtxx[xx[[1, i]], xx[[2, i]]][[All, j]].ISD.Qtxx[xx[[1, i]], xx[[2, i]]][[All, j]]);

(*cálculo de v(t/x)*)
For[j = 1, j <= m, j++, For[i = 1, i <= n, i++, vtxx[j, i] = 0] ]; Clear[Mvt];
For[j = 1, j <= m, j++, Mvt := Array[vtxx, {m, n}];
 For[i = 1, i <= n, i++, vtxx[j, i] = Exp[2*\[Beta].txx[[i]]]*(pp20[j] + pp27[j, i])]];
MatrixForm[Mvt]

(*****************************************************************************************)
(* 4. Cálculo de los Cuantiles e Intervalos de Confianza *)

Clear[wb, wbb, wb1, wb2, wb3, wb4, wb5, wb6, wb7, wb8, wb9, wb10, wb11, wb12, wb13, wb14,
wb15, wb16, wb17, wb18, wb19, wb20, wb21, wb22, wb23, wb24, wb25, wb26, wb27, wb28, wb29,
wb30, WBB, wbf, qt, qq, Cuantil, Gcuantil1, Gqtl1, Gcuantil2, Gqtl2, Mvtt0, LI, LS, MStxxx0,
pp30, \[Delta]2tx, v, LIC, LSC, LI, LS, SLi, SLs, VLi, VLs, IC0, TIC0, IC01, IC02, ic1, ic2,
fpi, fps, GIC, GSC, GO1, GO2];

(*cálculo de los omega (w)*)
(*Se ejecutan para c/muestra. Se almacenan para el cálculo de los cuantiles*)
MatrixForm[MAtxx]; MatrixForm[\[CapitalLambda]txx0]; MatrixForm[Mvt];
Clear[wb]; wb = (MAtxx - \[CapitalLambda]txx0)/((Mvt)^(1/2)); MatrixForm[wb]

(*Se calcula c/u de los t=12 wb producto de los valores de los "n" individuos*)
Clear[wbb]; wbb := wbb30; (*a la variable wbb se asigna en c/paso c/u de los wb
hallados en el paso anterior*)
Clear[wbbb, WB]; WB = Array[wbbb, m]; For[j = 1, j <= m, j++, wbbb[j] = Mean[wbb[[j, All]]]];
MatrixForm[WB]

(*Cálculo de los cuantiles C\[Alpha](t) y C(1-\[Alpha])(t) para las t=12 particiones
de tiempo*)
(*Se almacenan las matrices en una para el cálculo de los cuantiles*)
Clear[WBB, wbf]; WBB := {wb1, wb2, ..., wb30}; wbf := Transpose[WBB]; MatrixForm[wbf];
Clear[qt, qq, Cuantil]; qt = Array[qq, m];
For[j = 1, j <= m, j++, qq[j] = Quantile[wbf[[j, All]], {1/4, 3/4}]];
Export["qt.txt", qt, "Table"]; Dimensions[wbf]; MatrixForm[qt]

(*Gráfica de los cuantiles*)
Gcuantil1 := Array[Gqtl1, m]; For[j = 1, j <= m, j++, Gqtl1[j] = {tt[[j]], Cuantil[[j, 1]]}];
Gcuantil1
Gcuantil2 := Array[Gqtl2, m]; For[j = 1, j <= m, j++, Gqtl2[j] = {tt[[j]], Cuantil[[j, 2]]}];
Gcuantil2
ListLinePlot[{Gcuantil1, Gcuantil2}, PlotLegends -> {"LI", "LS"}, Mesh -> Full,
PlotMarkers -> Automatic, PlotRange -> {{20, 370}, {-1.5, 1.3}}, AxesOrigin -> {20, 0}]

(*Cálculo de los Intervalos de Confianza para la función de azar acumulada
(\[CapitalLambda]tx)*)
p = 0.5; Needs["HypothesisTesting`"]; (*p-esimo cuantil = 0.5(Mediana)*)
Clear[LI, LS, SLi, SLs];
(*Declaración de funciones*)
LI[j_, i_] := Lambda0[[j, i]] + Sqrt[Mvtt0[[j]][[i]]]*qt[[j]][[1]];
LS[j_, i_] := Lambda0[[j, i]] + Sqrt[Mvtt0[[j]][[i]]]*qt[[j]][[2]];
```

```
SLi[j_] := Sum[LI[j, i], {i, 1, n}]/n;
SLs[j_] := Sum[LS[j, i], {i, 1, n}]/n;

(*Se ejecutan las funciones y se almacenan*)
Clear[VLi, VLs]; VLi = Table[SLi[j], {j, 1, m}]; VLi; VLs = Table[SLs[j], {j, 1, m}]; VLs;
Clear[IC0, TIC0]; IC0 = {VLi, VLs}; TIC0 = Transpose[IC0]; MatrixForm[TIC0]
IC01 := Array[ic1, m]; IC02 := Array[ic2, m];
For[j = 1, j <= m, j++, ic1[j] = {tt0[[j]], VLi[[j]]}; ic2[j] = {tt0[[j]], VLs[[j]]}];
Export["IC01.txt", IC01, "Table"]; Export["IC02.txt", IC02, "Table"];

(*Gráfico: para una linea en general, el comportamiento del intervalo de confianza en el
tiempo*)
ListLinePlot[{IC01, IC02}, PlotLegends -> {"LI", "LS"}, Mesh -> Full,
PlotMarkers -> Automatic, AxesLabel -> {t, "\[CapitalLambda](t/x)"},
PlotRange -> {{20, 370}, {0, 1}}, AxesOrigin -> {20, 0}]
```

B.2. Gráficas adicionales

Aquí se muestra el código requerido para representar el valor de la función de azar acumulada a través de las dos covariables y fijando a la variable tiempo.

```
(*****************************************************************************************)
(*Declaración de funciones para los gráficos*)
v[j_, i_] := E^(2*\[Beta]0.txx0[[i, All]]) (pp2000[j] +
Qtxx0[xx0[[1, i]], xx0[[2, i]]][[All, j]].ISD0.Qtxx0[xx0[[1, i]], xx0[[2, i]]][[All, j]]);

(*Se crean funciones para los limites de confianza*)
LIC[j_, i_] := (Lambda0[[j, i]]) + Sqrt[v[j, i]]*Cuantil[[j, 1]];
LSC[j_, i_] := (Lambda0[[j, i]]) + Sqrt[v[j, i]]*Cuantil[[j, 2]];
fpi[j_, i_] := {x10[[i]], x20[[i]], LIC[j, i]}
fps[j_, i_] := {x10[[i]], x20[[i]], LSC[j, i]}

(*Se realiza la gráfica para c/tiempo t*)
GIC = Table[fpi[j, i], {j, 1, m}, {i, 1, n}];
GSC = Table[fps[j, i], {j, 1, m}, {i, 1, n}];
Export["GIC.txt", GIC, "Table"]; Export["GSC.txt", GSC, "Table"];

ListPlot3D[GIC, PlotStyle -> {Blue, Opacity[0.6]}, AxesLabel -> {"factura", "edad",
"\[CapitalLambda](t/x)"}, AxesOrigin -> {{20, 220}, {20, 80}, {0.1, 4.5}}];
ListPlot3D[GSC, PlotStyle -> {Orange, Opacity[0.4]}, Mesh -> None, AxesLabel -> {"factura",
"edad", "\[CapitalLambda](t/x)"},  AxesOrigin -> {{20, 220}, {20, 80}, {0.1, 4.5}}];
Show[%%, %]
```

Luego, para apreciar el comportamiento de los intervalos de confianza en el plano, fijamos a una de las covariables generando un corte sobre la gráfica, por ejemplo para un individuo con una edad de 30 años tenemos:

```
(*****************************************************************************************)
(*Se genera el corte para un individuo con una edad específica*)
plano = Graphics3D[Polygon[{{20, 30, 0.1}, {20, 30, 4.5}, {220, 30, 4.5}, {220, 30, 0.1}}]]

(*Gráfica que contrapone el plano de corte creado con la gráfica de los intervalos*)
Show[GIC, GSC, plano, ImageSize -> {600, 600}]
```

```
(*Muestra el corte realizado en el plano*)
Grafo = ListPlot3D[{GIC, GSC},
  AxesLabel -> {Style["factura", 14], Style["", 14],
    Style["\[CapitalLambda](t/x)", 14]},
  PlotRange -> {{20, 200}, {30, 30.05}, {0.4, 5.0}},
  ViewPoint -> {0, -Infinity, 0},
  PlotLabel -> Style["Edad: 30 años", 16, Bold], Mesh -> 445,
  MeshStyle -> Red, ImageSize -> {500, 500}]
```

Apéndice C

Algunos resultados numéricos

En esta sección presentamos algunos de los resultados obtenidos de aplicar los algoritmos dados en las Figuras 4.1, 4.2, 4.3 y 4.4. Así tenemos:

C.1. Cálculos para la muestra original

a. Cálculo de la función de riesgo base $\Lambda(t)$: Se muestra el valor de la función para cada uno de los doce tiempos definidos.

t	Λ(t)
1	0.01530
2	0.09971
3	0.15867
4	0.28264
5	0.34486
6	0.45085
7	0.48430
8	0.53916
9	0.57230
10	0.59267
11	0.61248
12	0.68442

Figura C.1: Función de riesgo base para la muestra original

b. Cálculo de la función de riesgo acumulada $\Lambda(t \mid x)$: Se muestra el valor de la función para cada uno de los doce tiempos definidos.

	individuo				
t	1	2	3	...	1535
1	0.01860	0.01903	0.01926	...	0.01362
2	0.12118	0.12404	0.12549		0.08879
3	0.19284	0.19738	0.19969		0.14129
4	0.34350	0.35158	0.35569		0.25168
5	0.41912	0.42898	0.43400		0.30709
6	0.54794	0.56083	0.56739	...	0.40147
7	0.58858	0.60243	0.60947		0.43125
8	0.65526	0.67067	0.67852		0.48010
9	0.69553	0.71189	0.72022		0.50961
10	0.72030	0.73724	0.74586		0.52776
11	0.74436	0.76187	0.77078		0.54539
12	0.83179	0.85136	0.86132	...	0.60945

Figura C.2: Función de riesgo acumulada para la muestra original

c. Cálculo de la segunda derivada $\ddot{l}\left(\widehat{\boldsymbol{\beta}}\right)$: Se muestra la matriz de varianza-covarianza producto del cálculo de la primera y segunda derivada.

$$\begin{pmatrix} 2.32868\times10^{-6} & 1.16699\times10^{-7} \\ 1.16699\times10^{-7} & 8.40509\times10^{-6} \end{pmatrix}$$

Figura C.3: Matriz de varianza-covarianza para la muestra original

d. Cálculo de la variable $\widehat{v}\left(t \mid \boldsymbol{x}\right)$: Se muestra el valor de la función para cada uno de los doce tiempos definidos.

	individuo				
t	1	2	3	...	1535
1	0.00001	0.00001	0.00001	...	0.00001
2	0.00012	0.00012	0.00012		0.00021
3	0.00023	0.00023	0.00023		0.00049
4	0.00055	0.00055	0.00056		0.00146
5	0.00077	0.00077	0.00078		0.00215
6	0.00121	0.00121	0.00123	...	0.00361
7	0.00138	0.00138	0.00139		0.00415
8	0.00167	0.00167	0.00169		0.00512
9	0.00186	0.00186	0.00188		0.00575
10	0.00199	0.00199	0.00201		0.00615
11	0.00211	0.00211	0.00213		0.00656
12	0.00261	0.00261	0.00263	...	0.00814

Figura C.4: Valor de la función $\widehat{v}\left(t \mid \boldsymbol{x}\right)$ para la muestra original

C.2. Cálculo para las muestras producto del muestreo bootstrap

Los siguientes son los parámetros estimados para las covariables de las diez primeras muestras de las cien obtenidas.

FACTURA	0.02224	0.01202	0.00762	0.01134	0.01227	0.01649	0.01328	0.00809	0.01945	0.00621
EDAD	-0.00680	-0.00525	-0.00529	-0.00771	-0.00663	-0.00792	-0.00503	-0.00768	-0.00704	-0.00682

Figura C.5: Parámetros obtenidos para las diez primeras muestras bootstrap

A continuación se presentan los cálculos obtenidos para una de las muestras, por ejemplo para la muestra B=100 tenemos:

a. Cálculo de la función de riesgo base $\Lambda\left(t\right)$: Se muestra el valor de la función para cada uno de los doce tiempos definidos.

b. Cálculo de la función de riesgo acumulada $\Lambda\left(t \mid x\right)$: Se muestra el valor de la función para cada uno de los doce tiempos definidos.

c. Cálculo de la segunda derivada $\ddot{l}\left(\widehat{\boldsymbol{\beta}}\right)$: Se muestra la matriz de varianza-covarianza producto del cálculo de la primera y segunda derivada.

d. Cálculo de la variable $\widehat{v}\left(t \mid \boldsymbol{x}\right)$: Se muestra el valor de la función para cada uno de los doce tiempos definidos.

t	Λ(t)
1	0.01709
2	0.12634
3	0.19890
4	0.35047
5	0.43234
6	0.55421
7	0.60530
8	0.65686
9	0.69586
10	0.71667
11	0.73662
12	0.81485

Figura C.6: Función de riesgo base para la muestra bootstrap B=100

	individuo				
t	1	2	3	...	1535
1	0.01765	0.01892	0.01650	...	0.01933
2	0.13046	0.13991	0.12201		0.14287
3	0.20538	0.22025	0.19208		0.22492
4	0.36190	0.38810	0.33846		0.39633
5	0.44644	0.47876	0.41752		0.48891
6	0.57229	0.61371	0.53521	...	0.62673
7	0.62505	0.67029	0.58455		0.68450
8	0.67829	0.72739	0.63434		0.74281
9	0.71856	0.77057	0.67200		0.78691
10	0.74005	0.79361	0.69210		0.81044
11	0.76065	0.81570	0.71136		0.83300
12	0.84143	0.90234	0.78692	...	0.92147

Figura C.7: Función de riesgo acumulada para la muestra bootstrap B=100

$$\begin{pmatrix} 3.35874\times10^{-6} & 1.65031\times10^{-7} \\ 1.65031\times10^{-7} & 8.40636\times10^{-6} \end{pmatrix}$$

Figura C.8: Matriz de varianza-covarianza para la muestra bootstrap B=100

	individuo				
t	1	2	3	...	1535
1	0.00001	0.00001	0.00001	...	0.00001
2	0.00009	0.00010	0.00010		0.00011
3	0.00015	0.00017	0.00020		0.00018
4	0.00029	0.00032	0.00046		0.00036
5	0.00038	0.00042	0.00065		0.00047
6	0.00054	0.00058	0.00098	...	0.00066
7	0.00061	0.00065	0.00113		0.00075
8	0.00068	0.00073	0.00131		0.00084
9	0.00074	0.00079	0.00144		0.00092
10	0.00078	0.00083	0.00152		0.00096
11	0.00081	0.00086	0.00159		0.00100
12	0.00094	0.00100	0.00190	...	0.00117

Figura C.9: Valor de la función $\widehat{v}\,(t \mid \boldsymbol{x})$ para la muestra bootstrap B=100

Bibliografía

Andersen, P. K. y Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study, *Ann Statist* **10**: 1100–1120.

Beran, R. y Hall, P. (1993). Interpolated nonparametric prediction intervals and confidence intervals, *Journal of the American Statistical Association* **55**: 643–652.

Bie, O., Borgan, O. y Liestol, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties., *Scandinavian Journal of Statistics* **14**: 221–233.

Borgan, O. y Liestol, K. (1990). A note on confidence interval and bands for the survival curves based on transformations., *Scandinavian Journal of Statistics* **18**: 35–41.

Breslow, N. E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**: 89–99.

Brookmeyer, R. y Crowley, J. (1982). A confidence interval for the median survival time, *Biometrics* **38**: 29–41.

Burr, D. y Doss, H. (1993). Confidence bands for the median survival time as a function of the covariates in the cox model, *Journal of the American Statistical Association* **88**: 1330–1340.

Chen, X. y Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles, *Ann Statist* **21**: 1166–1181.

Cochran, W. G. (1981). *Técnicas de Muestreo*, Compañía Editora Continental S.A.

Cox, D. R. (1972). Regression models and life-tables (with discussion), *Journal of the American Statistical Association* **34**: 187–220.

Cramér, H. (1928). On the composition of elementary errors, *Skand. Aktuarietidskr.* **11**: 13–74 and 141–180.

Dabrowska, D. y Doksum, K. (1987). Estimates and confidence intervals for median and mean life in the proportional hazard model, *Biometrika* **74**: 799–807.

Edgeworth, F. (1905). The law of the error, *Cambridge Philos Society* **20**: 36–66,113–141.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann Statist* **7**: 1–21.

Efron, B. y Tibshirani, R. J. (1993). *An introduction to the bootstrap*, Chapman & Hall.

Fleming, T. y Harrington, D. (1979). Nonparametric estimation of the survival distribution in censored data., *Unpublished manuscript* .

Gross, S. y Lai, T. L. (1996). Bootstrap methods for truncated and censored data, *Statist Sin* **6**: 509–530.

Gu, M. (1992). On the edgeworth expansion and bootstrap approximation for the cox regression model under random censorship, *Can J Statist* **20**: 399–414.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*, Springer.

Ho, Y. y Lee, M. (2005). Iterated smoothed bootstrap confidence intervals for population quantiles, *Ann Statist* **33**: 437–462.

Kalbfleisch, J. D. y Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, John Wiley & Sons.

Lai, T. L. y Wang, Q. (1993). Edgeworth expansions for symmetric statistics with applications to bootstrap methods, *Statist Sin* **3**: 517–542.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*, John Wiley & Sons.

Li, G., Hollander, M. y McKeague, I. Y. (1996). Nonparametric likelihood ratio confidence bands for quantile functions from incomplete data, *Ann Statist* **24**: 628–640.

Mathematica (2012). Version 9.0, *Wolfram Research* . Programa informático.

Miller, R. (1981). *Survival analysis*, Wiley.

Miller, R. y Halpern, J. (1982). Regression with censored data, *Biometrika* **69**: 521–531.

Nair, V. N. (1984). Confidence bands for survival functions with censored data: a comparative study, *Technometrics* **26**: 265–275.

Nelson, W. (1969). Hazard plotting for incomplete failure data, *Journal of Quality Technology* **1**: 27–52.

Organismo Supervisor de la Inversión Privada en Telecomunicaciones, O. (2012). Más acceso, mayor inclusión y desarrollo - informe de gestión 2007-2012, *Technical report*, Organismo Supervisor de la Inversión Privada en Telecomunicaciones, OSIPTEL. Consulta: 2013-03-01. Disponible en: http://www.osiptel.gob.pe/WebSiteAjax/Archivos/Publicaciones/Informe_de_Gestion_2007-2012.pdf.

Peterson, A. (1977). Expressing the kaplan-meier estimator as a function of empirical subsurvival functions., *Journal of the American Statistical Association* **72**: 47–50.

R (2012). Version 2.14.2, *The R Foundation for Statistical Computing* . Programa informático. Disponible en: http://cran.r-project.org/bin/windows/base/.

SPSS (2009). Version 18.0.0, *Polar Engineering and Consulting* . Programa informático.

Strawderman, R. L., Parzen, M. I. y Wells, M. T. (1997). Accurate confidence limits for quantiles under random censoring, *Biometrics* **53**: 1399–1415.

Survival (2013). Version 2.37-4, *Therneau, T. - Survival Analysis Library in R* . Consulta: 2013-03-27. Disponible en: http://cran.r-project.org/web/packages/survival/index.html.

Tsiatis, A. A. (1981). A large sample study of cox's regression model, *Ann Statist* **9**: 93–108.

Tze, L. L. y Zheng, S. (2006). Confidence intervals for suvival quantiles in the cox regression model, *Springer* **12**: 407–419.

Wallace, D. L. (1958). Asymptotic approximations to distributions, *Annals of mathematical statistics* **29**: 635–654.

Weisstein, E. W. (2013). Series edgeworth, *MathWorld - A Wolfram Web Resource* . Consulta: 2013-03-15. Disponible en: http://mathworld.wolfram.com/EdgeworthSeries.html.