PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE GRADUADOS

EL MODELO DE LARGA DURACIÓN WEIBULL-GEOMÉTRICA

TESIS PARA OPTAR POR EL GRADO DE MAGISTER EN ESTADÍSTICA

Presentado por:

Karina Hesi Torres Salinas

Asesor: Víctor Giancarlo Sal y Rosas Celi

Miembros del jurado:
Dr. Víctor Giancarlo Sal y Rosas Celi
Dra. Zaida Jesus Quiroz Cornejo
Dr. Cristian Luis Bayes Rodriguez

Lima, Julio 2018

Dedicatoria

Dedicado a quienes me dieron la oportunidad de cimentar los logros que voy obteniendo en la vida, Manuel y Margarita mis abuelos. Para los niños que me ayudan y enseñan que aún en la adversidad es posible sonreir y dar amor. Para mi ser perruno que partió hace poco. Han sido y son todos mi fuerza.

Agradecimentos

A Dios y María, por abrir ventanas en mi vida. A mi asesor, profesor Víctor Sal y Rosas por sus contribuciones, orientaciones, sugerencias y paciencia en el desarrollo de esta tesis. A mis profesores de la maestría Cristian Bayes y Zaida Quiroz, por sus orientaciones y recomendaciones y un agradecimiento especial al profesor José Flores por su contribución en la primera etapa de este trabajo. A mi familia, por su amor infinito. A mis amigos y compañeros del piso 14 del área de Calidad de Servicio ahora Experiencia Cliente, del área Gestión de Clientes y de Marketing por su infinito ánimo desde el inicio de la maestría, a Víctor por su gran paciencia, comprensión y ayuda precisa, a Amelia por hacerme siempre reir y apoyarme en las funciones del día a día, a Luchito por su gran chispa. A Iván y a Andrés por facilitarme los datos para la aplicación, se los agradezco infinitamente. A mis amigos de la vida y a los que conocí en la maestría, por ser mi barra. A mis mascotas por la alegría y amor que me dan.

Resumen

Los modelos de larga duración son una extensión de los modelos de supervivencia tradicional y nos permiten modelar una proporción de la población que no llegan a experimentar un evento de interés, incluso después de un largo periodo de seguimiento. En este trabajo se presenta y deduce la distribución de larga duración Weibull-Geométrica y su proceso de estimación e inferencia. Se desarrolló un estudio de simulación con el fin de evaluar el desempeño de las estimaciones y determinar si se recuperan los parámetros. Asimismo el modelo fue aplicado a una muestra de clientes que adquirieron y activaron una tarjeta de crédito entre enero a diciembre del año 2015 y donde el principal objetivo del analisis era entender el comportamiento del tiempo hasta la cancelación de la tarjeta de crédito del cliente. Comparamos al modelo de larga duración Weibull-Geométrica con otros modelos de larga duración, Exponencial-Geométrica y Weibull. Los resultados indican que nuestro modelo muestra un mejor ajuste en los datos.

Palabras-clave: distribución weibull, distribución geométrica, modelos de larga duración, esquema de la primera activación, modelos de mixtura, distribución weibull-geométrica.

Abstract

Long-term models are an extension of traditional survival models and allow us to model a proportion of the population that are immune to experience then event of interest. This thesis presents and discusses the Weibull-Geometric long-term distribution, its estimation and the inference procedure. A simulation study was developed in order to evaluate the performance of the estimates based on the bias and 95 % coverage. Finally, the model was also applied to a sample of customers who acquired and activated a credit card from January to December 2015. The aim of the analysis was to understand the behaviour of the time until the cancellation of the credit card and factors associated with no cancellation. We compared the Weibull-Geometric long-term model with the Exponential-Geometric and Weibull long-term models. The results indicate that our model shows a better fit in the data.
Keywords: Weibull distribution, long-term, mixture, Weibull-Geometric distribution.

Índice general

Lista de abreviaturas

v.a	Variable aleatoria .
fdp	Función de densidad de probabilidad .
fda	Función de distribución acumulada .
WGL	Modelo Weibull Geométrica de larga duración.
EGL	Modelo Exponencial Geométrica de larga duración.
WL	Modelo Weibull de larga duración.
MLE	Estimador de máxima verosimilitud.

Lista de símbolos

T	Tiempo hasta la ocurrencia del evento de interés.
f	Función de densidad.
F, G	Función de distribución acumulada.
S	Función de supervivencia.
h	Función de riesgo instantáneo.
δ	Indicador de censura

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones preliminares

En muchas ocasiones se desea estudiar el tiempo hasta que ocurra un evento de interés, denotado por T, y como éste podria estar asociado a ciertas características o variables. Este es el problema fundamental que se aborda en el análisis de supervivencia(**?**).

Sin embargo, es posible que exista un grupo de unidades de la población que no sean susceptibles. Matemáticamente, esto implicaria que

$$P(T = \infty) > 0$$

y, por lo tanto, que estemos ante un escenario de una variable aleatoria extendida **?**. Por ejemplo, tenemos el caso de pacientes con cáncer sometidos a resección curativa para metástasis hepáticas colorrectales, los datos se ajustaron a un modelo de cura sin mezcla para comparar la mortalidad después de la resección hepática con la mortalidad esperada para la población general combinada por sexo y edad (**??**). También se presentan modelos de fracción de cura para el área de calificación de riesgo crediticio donde una gran proporción del conjunto de datos no experimenta el evento de interés referida al inclumplimiento en la cartera de préstamos personales Tong et al. (2012).

En esta tesis se presenta un modelo donde se considera que una fracción de la población no es susceptible al evento de interés.

1.2. Objetivos

El objetivo general de la tesis es estudiar el modelo de larga duración Weibull-Geométrico. De manera específica:

- Revisar conceptos preliminares de la distribución Weibull-Geométrica, datos censurados y modelos de mixtura.
- Presentar y estudiar el modelo de larga duración Weibull-Geométrica.
- Realizaremos un estudio de simulación para evaluar el desempeño de los estimadores del modelo, para muestras pequeñas, obtenidos via el método de máxima verosimilitud.
- Aplicar el modelo a una muestra de clientes que adquirieron y activaron una tarjeta de crédito durante el periodo de enero a diciembre del año 2015 y donde el evento

de interés esta definido como la cancelación de la tarjeta de crédito adquirida por un cliente.

1.3. Organización del trabajo

En el Capítulo 2, presentamos conceptos preliminares. En el Capítulo 3, se realizará la presentación y deducción del modelo de larga duración Weibull-Geométrica, se discutirá el proceso de estimación e inferencia via máxima verosimilitud. En el Capítulo 4, se presenta un pequeño estudio se simulación para evaluar el desempeño de los estimadores propuestos para muestras pequeñas. En el Capítulo 5 se presenta los resultados del análisis de una muestra de clientes que han adquirido una tarjeta de crédito durante el año 2015 a los cuales se les hizo un seguimiento de 24 meses con el fin de estudiar el tiempo hasta la cancelación de la tarjeta de crédito. Finalmente, en el Capítulo 6 discutimos algunas conclusiones y sugerencias para trabajos futuros.

Capítulo 2

Conceptos preliminares

En el análisis de supervivencia la variable de interes es el tiempo, denotado por T, hasta la ocurrencia de cierto evento de interés en la población. La distribución de la variable aleatoria (v.a) T se puede describir de varias formas:

Función de distribución acumulada

Es la probabilidad de que a un individuo de la población, seleccionado al azar, le ocurra el evento antes de o en el tiempo t, se denota por f.d.a y es dada por:

$$F(t) = P(T \leq t) = \int_0^t f(u)du,\ t > 0$$

donde $f(\cdot)$ es la función de densidad de T

Las propiedades que deben cumplir las funciones de probabilidad acumulada

i) F es una función no decreciente

ii) La función tiende a cero por la izquierda. Es decir

$$\lim_{h \to -\infty} F(h) = 0$$

iii) Si nos acercamos al infinito (derecha), la función tiende a 1. Es decir

$$\lim_{h \to \infty} F(h) = 1$$

Si T es una variable continua, entonces la relación entre la función de densidad $f(\cdot)$ y la f.d.a $F(\cdot)$, está dada por:

$$f(t) \quad = \quad \frac{\partial F(t)}{\partial t}, \qquad F(t) = \int_0^t f(u)du$$

Función de supervivencia

Es la probabilidad de que un individuo de la población, seleccionado al azar, sobreviva hasta el tiempo t:

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)du,\ t > 0$$

Función de riesgo

Es el riesgo instantáneo de que el evento ocurra en el intervalo $[t, t+\triangle t]$, dado que no ha ocurrido hasta el tiempo t. Es decir, la posibilidad de que un individuo que ha sobrevivido hasta el tiempo t, le ocurra el evento en el siguiente instante en el tiempo. Esta función se define como:

$$h(t) = \lim_{\triangle t \to 0^+} \frac{P(t \leq T \leq t+\triangle t \mid T \geq t)}{\triangle t} = \frac{f(t)}{S(t)}$$

Note que se debe cumplir que

i) $h(t) \geq 0, \ \forall t$

ii) $h(\cdot)$ no tiene límite superior

2.1. Distribución Weibull

Una variable aleatoria continua no negativa, denotada por T, se dice que sigue una distribución Weibull con parámetros de forma y escala λ_1 y λ_2 respectivamente, denotada por $T \sim Weibull(\lambda_1, \lambda_2)$, si su función de densidad esta dada por

$$f(t) = \frac{\lambda_1}{\lambda_2}\left(\frac{t}{\lambda_2}\right)^{\lambda_1 - 1} e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}, \ t > 0 \tag{2.1}$$

Las funciones de probabilidad acumulada, supervivencia y riesgo instantáneo están dadas por:

$$\begin{aligned} F(t) &= 1 - e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}} \\ S(t) &= e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}} \\ h(t) &= \frac{\lambda_1}{\lambda_2}\left(\frac{t}{\lambda_2}\right)^{\lambda_1 - 1} \end{aligned}$$

para $\lambda_1, \lambda_2 > 0 \quad$ y $\quad t > 0$.

Note que, cuando $\lambda_1 = 1$, T sigue una distribución Exponencial con parámetro λ y denotada por $T \sim Exp(\lambda)$. En este caso particular se tiene:

$$\begin{aligned} f(t) &= \lambda e^{(-\lambda t)} \\ F(t) &= 1 - e^{-\lambda t} \\ S(t) &= e^{-\lambda t} \\ h(t) &= \lambda \end{aligned}$$

para $\lambda > 0 \quad$ y $\quad t > 0$.

2.2. Distribución Weibull-Geométrica

La distribución Weibull es una de las distribuciones más usadas para modelar datos de tiempos de vida, es flexible y se utiliza para modelar diversos datos de tiempos de vida

con función de riesgo monótona, sin embargo no es adecuada para modelar casos en que la función de riesgo tiene forma unimodal o de bañera, los cuales son comunes en los estudios de supervivencia, por ejemplo en estudios biológicos y de fiabilidad (Barreto-Souza et al., 2011). En la literatura se han propuesto varias distribuciones para extender la distribución de Weibull, con el fin de modelar funciones de riesgos con forma unimodal o de bañera. Entre ellas tenemos la distribución Exponencial-Geométrica (EG) introducida por Adamidis and Loukas (1998) y la distribución Exponencial-Geométrica Extendida (EEG) propuesta por Adamidis et al. (2005). Marshall and Olkin (1997) introdujeron un método para agregar un nuevo parámetro a una familia de distribuciones (aplicadas a las familias exponencial y weibull).

Barreto-Souza et al. (2011) presentaron la distribución Weibull-Geométrica (WG) obtenida a partir de la composición de las distribuciones Weibull y Geométrica, bajo un esquema de factores de riesgos latentes. Esta distribución contiene las distribuciones EEG, EG y Weibull como modelos particulares. La función de riesgo de la distribución WG puede tomar formas más generales, es útil para modelar datos con tasa de fallo unimodal.

El modelo

Sea M una variable aleatoria discreta que describe la cantidad de factores de riesgos latentes relacionados con la ocurrencia de un evento de interés, para un individuo de la población. Asumimos que M tiene una distribución geométrica con parámetro p, denotado por $M \sim Geo(p)$, cuya función de probabilidad está dada por:

$$P(M = m) = p(1-p)^{m-1}, \quad m \in 1, 2, ... \tag{2.2}$$

donde $0 < p < 1$. Los tiempos de activación de estos factores de riesgos corresponden a las variables aleatorias contínuas no negativas denotado por $Z_i,\ i = 1, 2, ..., M$ independientes entre sí e independientes de M y con función de distribución acumulada $G(\cdot)$. El tiempo hasta la ocurrencia del evento es definido por la variable aleatoria $T = Z_{(R)}$, donde, $Z_{(1)} \leq ... \leq Z_{(R)} \leq ... \leq Z_{(M)}$ son los estadísticos de orden de Z_i, tal que R depende de M, y donde R puede ser una constante que depende de M o incluso puede ser una variable aleatoria especificada mediante una distribución condicional para R dado $M \geq 1$.

Con las condiciones descritas anteriormente, la función de distribución condicional de T dado $M = m$ y $R = r$ está dado por:

$$\begin{aligned} F_{T|m,r}(t) &= P[T \leq t \mid M = m, R = r] \\ &= \sum_{i=r}^{m} \binom{m}{i} [G(t)]^i [1 - G(t)]^{m-i} \end{aligned} \tag{2.3}$$

Este resultado se demuestra en el apéndice A.1. Asumimos que R es una variable aleatoria, la distribución condicional de R dado M es una distribución uniforme discreta en $1, 2, ..., m$ (esquema de activación aleatoria) con probabilidad $1/m$, verificamos que la distribución mar-

ginal de T está dada por:

$$\begin{aligned} F(t) &= \sum_{m=1}^{\infty}\sum_{r=1}^{m} P(T \leq t, M = m, R = r) \\ &= \sum_{m=1}^{\infty}\sum_{r=1}^{m} P[Z_{(R)} \leq t \mid M = m, R = r]P[R = r \mid M = m]P[M = m] \\ &= 1 - \sum_{m=1}^{\infty}\left(\sum_{r=0}^{m}(m-r)B(r, m, G(t))\right)\frac{1}{m}P[M = m] \\ &= 1 - (1 - G(t))\sum_{m=1}^{\infty} P[M = m] = G(t) \end{aligned} \tag{2.4}$$

donde $B(x, m, G(t))$ es la fdp de la distribución binomial, con parámetros m y $G(t)$ y $P[M = m]$ está dado en (2.2). Observar que la distribución marginal de T dada en (2.4) es la misma que la distribución de las variables aleatorias $Z_i's$

Ahora para el caso donde $R = r$ es fijo, entonces la distribución marginal de T está dada por:

$$\begin{aligned} F(t) &= \sum_{m=1}^{\infty} IB(F(t); r; m - r + 1)P[M = m] \\ &= \sum_{m=1}^{\infty} m \begin{pmatrix} m-1 \\ r-1 \end{pmatrix} \int_0^{G(t)} u^{r-1}(1-u)^{m-r} du \; p_m \end{aligned} \tag{2.5}$$

donde $IB(x, a, b)$ es la función beta incompleta y $p_m = P[M = m]$ está dado en (2.2)

Ahora supongamos que el evento de interés ocurre debido a que cualquiera de los posibles factores de riesgos latentes se activa, pero para $R = 1$, esto es, $T = \text{mín}\{Z_1, ..., Z_M\}$. Este caso corresponde al esquema de la primera activación Cooner et al. (2007), reemplazamos $r = 1$ en (2.5) entonces, la distribución marginal de T en este caso, esta dada por:

$$F(t) = \frac{G(t)}{1 - (1-p)(1-G(t))}, \tag{2.6}$$

la función de supervivencia está dada por:

$$S(t) = \frac{p(1-G(t))}{1 - (1-p)(1-G(t))} \tag{2.7}$$

y la función de riesgo:

$$h(t) = \frac{g(t)}{(1-G(t))(1-(1-p)(1-G(t)))} \tag{2.8}$$

Se puede elegir diferentes distribuciones $G(t)$ para las variables aleatorias $Z_i's$. Para el modelo presentado por Barreto-Souza et al. (2011), las variables aleatorias $Z_i's$ siguen una distribución Weibull con parámetros λ_1 y λ_2 dada en (2.1); si reemplazamos en (2.7) y (2.8)

obtenemos la fda y la función de supervivencia de la distribución Weibull-Geométrica bajo el esquema de la primera activación.

Distribución Weibull-Geométrica bajo el esquema de la primera activación

Sea T una variable aleatoria no negativa, se dice que sigue una distribución Weibull-Geometrica con parámetros, p, λ_1 y λ_2, denotada por $WG(p, \lambda_1, \lambda_2)$, si su función de densidad está dada por:

$$f_{WG}(t) = p\frac{\lambda_1 t^{\lambda_1-1}}{\lambda_2^{\lambda_1}} e^{-(\frac{t}{\lambda_2})^{\lambda_1}} \left\{1-(1-p)e^{-(\frac{t}{\lambda_2})^{\lambda_1}}\right\}^{-2} \tag{2.9}$$

La función de distribución acumulada está dada por:

$$F_{WG}(t) = \frac{1-e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}}{1-(1-p)e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}} \tag{2.10}$$

Adicionalmente, las funciones de supervivencia y de riesgo instantáneo están dadas por:

$$S_{WG}(t) = \frac{pe^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}}{1-(1-p)e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}} \tag{2.11}$$

$$h_{WG}(t) = \frac{\lambda_1\lambda_2^{-\lambda_1}t^{\lambda_1-1}}{1-(1-p)e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}} \tag{2.12}$$

Asimismo se obtienen sub-modelos particulares. En particular, cuando $\lambda_1 = 1$ obtenemos la distribución Exponencial-Geométrica (Adamidis and Loukas, 1998), cuando $\lambda_1 = 1$ y para cualquier $p < 1$, se obtiene la distribución Exponencial-Geométrica Extendida (Adamidis et al., 2005). Finalmente, si $p = 0$ obtenemos la distribución Weibull (2.1).

Cuantiles

El cuantil $q-$ésimo de la distribución Weibull-Geométrica se obtiene invirtiendo la función acumulada de probabilidad dada en (2.10) y es sencillo probar que está dado por:

$$t_q = \lambda_2\left\{log\left(\frac{1-(1-p)q}{1-q}\right)\right\}^{\lambda_1} \tag{2.13}$$

Momentos

El r-ésimo momento de T está dado por:

$$E(T^r) = p\lambda_2^{-r}\Gamma(\lambda_1 r+1)\Phi(p, r\lambda_1, 1) \tag{2.14}$$

donde

$$\Phi(z, s, a) = \frac{1}{\Gamma(s)}\int_0^\infty t^{s-1}e^{-at}(1-ze^{-t})^{-1}dt$$

para $z < 1$, $a > 0$ y $s > 0$ es la función Lerch trascendente.

2.3. Datos Censurados y verosimilitud

En algunos casos, cuando los datos son recolectados durante un periodo de tiempo determinado, ocurra que el evento no podrá ser observado en algunos individuos y solo se registra información parcial sobre el tiempo de interés. A este resultado se le conoce como dato censurado.

Se pueden presentar tres tipos de datos censurados: por la derecha, cuando solo se conoce que el evento no ha ocurrido hasta el tiempo $T = t$; por la izquierda, cuando se sabe que el evento ha ocurrido antes del inicio del seguimiento del individuo y censura por intervalo, cuando se sabe que el evento ocurrió entre dos puntos en el tiempo.

En esta tesis el tipo de censura a abordar es censura por la derecha. Sea T el tiempo hasta el evento de interés con una distribución acumulada de probabilidad denotada por F y sea Y denominado tiempo de censura con función de distribución acumulada G. Entonces, en el caso de censura por la derecha, la información del tiempo esta dada por $(\tilde{T}, \Delta)$ donde

$$\tilde{T} = \text{mín}\,\{\,T, Y\,\}, \quad \Delta = \begin{cases} 1, & \text{si } T \leq Y \\ 0, & \text{si } T > Y \end{cases}$$

donde $\tilde{T}$ es el tiempo observado y Δ es el indicador de censura.

En este trabajo, los datos censurados se estudian por la derecha, entonces es necesario tener en cuenta que tendremos dos tipos de información, la información de los individuos que experimentan el evento y por tanto se puede medir el tiempo y la información de los individuos censurados, cuyos tiempos no se han podido medir o son parciales, en estos individuos el evento no ha ocurrido o se dará después del periodo de estudio; por tipo de información la contribución en la verosimilitud será diferente.

Cuando el evento ocurre en el individuo, estoy observando la variable en estudio, por lo tanto su contribución a la verosimilitud es la función de densidad $f(\cdot)$, pero cuando existe una observación o individuo censurado por la derecha, lo que se tiene hasta ese instante es la función de supervivencia $S(\cdot)$ y eso se toma en cuenta en la verosimilitud.

Consideremos una muestra de n observaciones $(\tilde{t}_1, \delta_1), \ldots, (\tilde{t}_n, \delta_n)$. La verosimilitud para un conjunto de datos en estudio denotados por, asumiendo que T es independiente de G, se define como:

$$L(\theta) \quad \propto \quad \prod_{i=1}^{n} f(\tilde{t}_i)^{\delta_i} S(\tilde{t}_i)^{1-\delta_i} \tag{2.15}$$

Donde n es el número de observaciones en la muestra y f y F son la función de densidad y distribución acumulada de T, respectivamente

2.4. Estimador de Kaplan Meier

Conocido como el estimador del límite del producto, está basado en la descomposición de la función de supervivencia en un producto de probabilidades condicionadas.

Sea $t_1 < t_2 < ... < t_k$, los tiempos donde al menos un evento es observado, en base a estos tiempos, particionamos la muestra para formar $k + 1$ intervalos con los tiempos donde el evento ha ocurrido:

$$[t_0, t_1),\ [t_0, t_1)\ \ldots\ [t_k, t_{k+1})$$

donde $t_0 = 0$ y $t_{k+1} = \infty$. Definimos:

$d_j = \#$ de individuos que experimentan el evento en el tiempo t_j

$m_j = \#$ de individuos censurados en el intervalo $[t_j, t_{j+1})$

$n_j = \#$ de individuos en riesgo (o van quedando) en el instante previo a t_j. Es decir

$$n_j = (m_j + d_j) + (m_{j+1} + d_{j+1}) + (m_k + d_k)$$

De manera especifica, dentro de cada intervalo formado $[t_j, t_{j+1})$, tenemos d_j eventos y m_j datos censurados que ocurrieron en los tiempos t_i y $t_{j1} < t_{j2} < \ldots < t_{jm_j}$, respectivamente.

Kaplan and Meier (1958) propusieron el estimador puntual de la función de supervivencia y esta definido por

$$\widehat{S}^{KM}(t) = \prod_{j|t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

2.5. Modelos de larga duración vía mixtura

En el análisis de supervivencia tradicional se asume que el evento de interés le va a ocurrir a todos los individuos o personas de una población en algún momento, aunque sea en el infinito. Por tanto, la función de supervivencia $S(\cdot)$ tiende a cero a medida que el tiempo aumenta.

En contraste, en los modelos de larga duración se considera que existe una fracción de unidades de observación o individuos inmunes en los que no se llegará a observar el evento de interés, incluso después de un largo periodo de seguimiento, este grupo de individuos se denomina en la literatura como no susceptibles o curado (sobrevivientes a largo plazo), mientras que el grupo de individuos que experimentan el evento se denominan susceptibles.

Los modelos de larga duración o llamados también modelos de fracción de cura se han estudiado ampliamente en el ámbito de las ciencias médicas y de la salud, debido a un importante avance en la mejora de los tratamientos respecto a una variedad de enfermedades, pero también debido a que en muchos fenómenos sociales algunos individuos no son susceptibles al evento de interés. Estas aplicaciones se pueden encontrar en estudios biomédicos, finanzas, criminología, ingeniería, entre otras áreas (Sposto, 2002; Tong et al., 2012)

Bajo este contexto, el tiempo T hasta la ocurrencia del evento de interés se describe como:

$$T = \begin{cases} \infty \\ \mathrm{T}_s \end{cases}$$

donde $T_s \sim F_s$ es una variable aleatoria que mide el tiempo de ocurrencia al evento entre las unidades susceptibles. En el modelo de larga duración el tiempo T se denomina variable aleatoria extendida.

El enfoque clásico para el estudio de datos de supervivencia de larga duración es el modelo de mixtura propuesto por Boag (1949). En este modelo se asume que la población de unidades

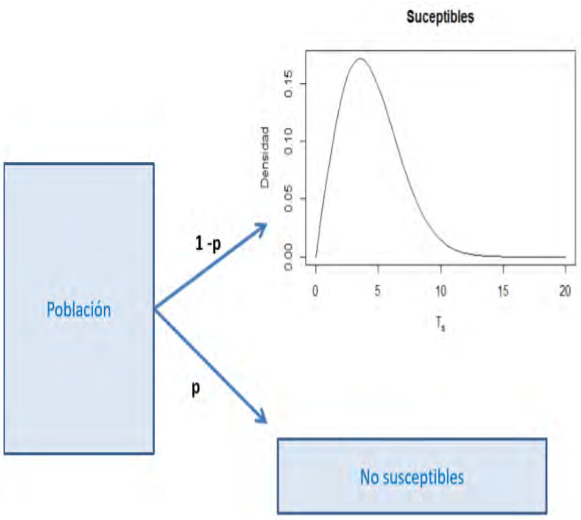

Figura 2.1: Modelo de mixtura

de estudio está formada por unidades inmunes y susceptibles al evento de interés, Boag refiere que el modelo de mixtura es de la forma descrita en la figura 2.1.
Este modelo ha sido estudiado intensamente por muchos autores, por ejemplo en Berkson and Gage (1952). Bajo este modelo la función de supervivencia de la población, denotada por $S_p(\cdot)$, está dada por:

$$S_p(t) \quad = \quad p_0 S_\infty(t) + (1 - p_0) S^*(t) \tag{2.16}$$

donde $S_\infty(\cdot)$ es la supervivencia al infinito (población de los no susceptibles), en esta población todo está concentrado en un solo valor, el infinito, por tanto la función de probabilidad de esta población corresponde a una variable discreta que solamente asume el valor infinito, entonces:

$$S_\infty(t) = P(\infty > t) = 1$$

Dado que una variable aleatoria no puede asumir el valor infinito, entonces a esta variable T se le llama extendida, porque la realidad nos obliga a considerar el infinito, ya que sin esta consideración no podríamos trabajar en el modelo de larga duración. Por tanto, se extiende la definición de variable aleatoria, ya que ésta asume valores reales y el infinito no es un número real, por eso se denomina a T variable aleatoria extendida. La función dada en (2.16) tiene las siguientes propiedades:

i) $S_p(0) = 1$

ii) $S_p(t)$ es decreciente

iii) $\lim_{t\to\infty} S_p(t) = p_0$

El modelo dado en (2.16) puede ser reescrito como:

$$S_p(t) \quad = \quad p_0 + (1 - p_0) S^*(t) \tag{2.17}$$

donde $S^*(\cdot)$ representa la función de supervivencia de los individuos no curados (o más generalmente, de las unidades susceptibles al evento de interés) y $p_0 = \lim_{t\to\infty} S(t)$ representa la fracción de curados o no suceptibles al evento de interés (Boag, 1949).

Para $S^*(\cdot)$ se puede considerar cualquier distribución de supervivencia, entre las más comunes están la distribución exponencial, weibull, gompertz y la distribución log-normal Yin and Ibrahim (2005). En otros artículos se han considerado también modelos con diferentes suposiciones sobre la distribución de tiempos de vida de los individuos susceptibles $S^*(\cdot)$, para el caso paramétrico (revisar Farewell (1982); Goldman (1984); Ghitany et al. (1995)).

Para modelar la proporción de unidades que no son susceptibles al evento de interés p_0, Farewell (1977) consideró un modelo de regresión logística:

$$p_0 = \frac{exp(X^T\beta)}{1 + exp(X^T\beta)} \tag{2.18}$$

donde X y β son los vectores de las covariables y los coeficientes de regresión respectivamente.

Farewell (1986) indica algunas restricciones que uno enfrenta al aplicar los modelos de larga duración, una de ellas es que se necesita una fuerte evidencia para la existencia de dos o más subestructuras en la población al aplicar los modelos de larga duración en un conjunto de datos. La presencia de una proporción inmune al evento (o fracción de cura) de un conjunto de datos, se puede visualizar cuando la curva de supervivencia de Kaplan-Meier alcanza un nivel estable, la supervivencia va tener una asíntota en p_0.

Cooner et al. (2007) propuso una clase de modelos jerárquicos flexibles con riesgos o factores competitivos latentes y diferentes esquemas de activación, donde el tiempo de vida asociado con un riesgo particular no es observable, sino que solo se observa el valor mínimo, máximo o un valor aleatorio.

Los modelos de fracción de cura basados en sistemas de activación latentes suponen que un evento de interés ocurre debido a M factores de riesgo latentes. Cooner et al. (2007) indica que se requieren R de M factores latentes para que se produzca el evento de interés. En muchos procesos biológicos, R indica los factores de resistencia del sistema inmune del individuo (Cancho et al. (2011)), entonces R puede ser una constante fija, una función de M o incluso una variable aleatoria especificada a través de una distribución condicional para R dado M.

2.6. Modelo de larga duración bajo el esquema de la primera activación

En los modelos de larga duración, el esquema de la primera activación supone que el evento de interés ocurre debido a que se activa cualquiera de los M posibles factores de riesgo latente; en este caso el evento ocurre para $R = 1$, el tiempo T hasta la ocurrencia del evento de interés se describe como:

$$T = \begin{cases} \infty, & \text{si } M = 0, \\ \text{mín}\{Z_1, ..., Z_M\}, & \text{si } M \geq 1 \end{cases}$$

La función de supervivencia de la población con el esquema de la primera activación y $M \sim Geo(p)$ está dada por:

$$\begin{aligned}
S_p(t) &= P(T > t \cap M = 0) + P(T > t \cap M = 1) + P(T > t \cap M = 2) + ... \\
&= P(T > t \cap M = 0) + \sum_{m=1}^{\infty} P(T > t \cap M = m) \\
&= P(M = 0)P(T > t \mid M = 0) + \sum_{m=1}^{\infty} P(M = m)P(T > t \mid M = m)) \\
&= (p)P(\infty > t \mid M = 0) + \sum_{m}^{\infty} p(1-p)^m P(\text{mín}\{Z_1, ...Z_M\} > t \mid M = m) \\
&= (p)(1) + \sum_{m=1}^{\infty} p(1-p)^m P(\text{mín}\{Z_1, ...Z_m\} > t \mid M = m) \\
&= p + \sum_{m=1}^{\infty} p(1-p)^m [1 - F(t)]^m \\
&= p + \sum_{m=1}^{\infty} [S(t)]^m = p + p\sum_{m=1}^{\infty} [S(t)(1-p)]^m \\
&= \frac{p}{1-(1-p)S(t)}
\end{aligned} \tag{2.19}$$

La fración de cura es $p_0 = p$. La función de densidad de la población está dada por:

$$f_p(t) = p(1-p)f(t)[1-(1-p)S(t)]^{-2} \tag{2.20}$$

y la función de riesgo de la población está dado por:

$$h_p(t) = (1-p)f(t)[1-(1-p)S(t)]^{-1} \tag{2.21}$$

Otro ejemplo especial de este esquema son los modelos YCIS (formulada por Yakovlev et al. (1993); Yakovlev (1996); Yakovlev and Tsodikov (1996); Chen et al. (1999), modelos de tumores estocásticos), donde M se distribuye como una Poisson, $M \sim Poisson(p)$, la función de supervivencia de la población está dada por: $S_p(t) = e^{-p(1-S(t))}$, con fracción de cura: e^{-p}
En los modelos de tipo Berkson and Gage (1952), M es binario, con solo un evento latente, $M \sim Bernoulli(p)$, donde p es la probabilidad de activación, la función de supervivencia poblacional está dada por: $S_p(t) = 1 - p(1 - S(t))$, con fracción de cura: $1 - p$
Se pueden considerar diferentes opciones para la distribución de las variables latentes $Z_i'^s$, con las cuales se pueden obtener nuevas familias de distribución. Para el caso de la distribución de Exponencial-Poisson propuesta por Kuş (2007) la variable Z_i sigue una distribución exponencial bajo el esquema de la primera activación.

Capítulo 3

Modelo de larga duración Weibull-Geométrica

3.1. Formulación del modelo

Supongamos que para un individuo de la población, sea M una variable aleatoria discreta que describe los factores de riesgos latente de los tiempos de activación de estos. Asumiremos que esta variable aleatoria tiene una distribución geométrica con parámetro p:

$$M \sim Geo(p)$$

La función densidad de probabilidad de la distribución geométrica está dada por:

$$P(M = m) = p(1-p)^m, \quad m \in 0, 1, 2, ... \tag{3.1}$$

donde $0 < p < 1$. Los tiempos de activación de estos factores corresponden a las variables aleatorias $Z_i, i \in N^+$, independientes entre sí e independientes de M y con distribución Weibull con parámetros λ_1 y λ_2, esto es

$$Z_i \sim Weibull(\lambda_1, \lambda_2)$$

El tiempo, T, hasta la ocurrencia del evento de interés corresponde al tiempo de la primera activación (Cooner et al. (2007)), está definida por:

$$T = \begin{cases} \infty, & \text{si} \quad M = 0, \\ \text{mínimo}\{ Z_1, \ldots, Z_M \}, & \text{si} \quad M \geq 1. \end{cases} \tag{3.2}$$

Se puede verificar que T dado en (3.2) genera un modelo de larga duración como los descrito en el capítulo anterior. Mas especificamente, se tiene

$$\begin{aligned}
S_p(t) &= P(T>t\cap M=0)+P(T>t\cap M=1)+P(T>t\cap M=2)+... \\
&= P(T>t\cap M=0)+\sum_{m=1}^{\infty}P(T>t\cap M=m) \\
&= P(M=0)P(T>t\mid M=0)+\sum_{m=1}^{\infty}P(M=m)P(T>t\mid M=m)) \\
&= (p)P(\infty>t\mid M=0)+\sum_{m=1}^{\infty}p(1-p)^m P(\text{mín}\,\{Z_1,...Z_M\}>t\mid M=m) \\
&= (p)(1)+\sum_{m=1}^{\infty}p(1-p)^m P(\text{mín}\,\{Z_1,...Z_m\}>t\mid M=m)
\end{aligned}$$

Dado que $Z_i\sim Weibull(\lambda_1,\lambda_2)$, entonces el mín $\{Z_1,...Z_m\}\sim Weibull(m\lambda_1,\lambda_2)$. Entonces

$$\begin{aligned}
S_p(t) &= (p)(1)+\sum_{m=1}^{\infty}p(1-p)^m(1-(1-e^{-\left(\frac{t}{\lambda_2}\right)^{m\lambda_1}})) \\
&= p+\sum_{m=1}^{\infty}p(1-p)^m e^{-\left(\frac{t}{\lambda_2}\right)^{m\lambda_1}}
\end{aligned}$$

Si definimos $\gamma=e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}$, la ecuación anterior se reduce a

$$S_p(t)=p+\sum_{m=1}^{\infty}p(1-p)^m\gamma^m=p+p\sum_{m=1}^{\infty}((1-p)\gamma)^m \tag{3.3}$$

Como $0<p<1,\quad 0<\gamma<1$. Es sencillo notar que bajo estas condiciones la serie es convergente y se reduce a:

$$\begin{aligned}
S_p(t) &= p+p\frac{(1-p)\gamma}{1-(1-p)\gamma} \\
&= p+(1-p)\frac{p\gamma}{1-(1-p)\gamma}
\end{aligned} \tag{3.4}$$

Entonces, la función de supervivencia de las unidades susceptibles al evento esta dado por:

$$S(t) = \frac{p\gamma}{1-(1-p)\gamma}=\frac{pe^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}}{1-(1-p)e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}} \tag{3.5}$$

Reemplazamos $S(\cdot)$ en (3.6) para obtener la función de supervivencia del modelo de larga duración dada en (2.15)

$$S_p(t) = p+(1-p)\frac{pe^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}}{1-(1-p)e^{-\left(\frac{t}{\lambda_2}\right)^{\lambda_1}}}=\frac{p}{1-(1-p)e^{-(\frac{t}{\lambda_2})^{\lambda_1}}} \tag{3.6}$$

Note que la fracción de cura, $p_0=P(M=0)=p$, entonces reemplazando $p=p_0$, en (3.8),

tenemos que:

$$S_p(t) = \frac{p_0}{1-(1-p_0)e^{-(\frac{t}{\lambda_2})^{\lambda_1}}} \tag{3.7}$$

Podemos notar que tenemos submodelos particulares de del modelo Weibull-Geométrico de larga duración. En particular, cuando $p = 1$ se reemplaza en la ecuación (3.7) la función de supervivencia de las unidades susceptibles al evento de interés $S(t)$ se distribuye como una Weibull, reemplazandola en (2.15), se obtiene el modelo Weibull de larga duración, dado por:

$$S(t) = p_0 + (1-p_0)e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} \tag{3.8}$$

La función $S(t)$ se distribuye como una Exponencial-Geométrica (EG) cuando $\lambda_1 = 1$. Entonces, reemplazándola en $S_p(t)$, y colocando todo en función de p_0, de la función dada en (3.9), se obtiene el modelo Exponencial-Geométrico de larga duración:

$$S_p(t) = \frac{p_0}{1-(1-p_0)e^{-(\frac{t}{\lambda_2})}} \tag{3.9}$$

La función de densidad (extendida) de T, esta dada por

$$f(t) = -\frac{d}{dt}S(t) = p_0(1-p_0)\frac{\lambda_1 t^{\lambda_1-1}}{\lambda_2^{\lambda_1}}e^{-(\frac{t}{\lambda_2})^{\lambda_1}}\left\{1-(1-p_0)e^{-(\frac{t}{\lambda_2})^{\lambda_1}}\right\}^{-2} \tag{3.10}$$

Un modelo de supervivencia similar Exponencial-Poisson obtenido fue estudiado en Yakovlev y Tsodikov (1993), Tsodikov (1998) y Chen et al. (1999)

Adicionalmente, deseamos ahora estudiar el efecto de k-covariables, $X = (X_1, \ldots, X_k)$, las cuales se relacionarán con la proporción de unidades inmunes al evento de interés a través de la regresión logística propuesta en Farewell (1977)

$$p_0 = \frac{exp(X^T\beta)}{1+exp(X^T\beta)} \tag{3.11}$$

donde $\beta = (\beta_1, ..., \beta_k)^T$ es el vector de coeficientes asociado a las covariables. De este modo el vector de los parámetros del modelo será $\theta = (\lambda_1, \lambda_2, \beta) \in \ominus = R^+ x R^k$

3.2. Estructura de datos y verosimilitud

El tipo de censura a abordar en esta tesis es censura por la derecha. Esta censura se da cuando para un individuo el tiempo T hasta el evento de interés no es observado pero se sabe que es mayor a un tiempo C (denominado tiempo de censura). Entonces, en el caso de censura por la derecha, la información del tiempo esta dada por $(\tilde{T}, \Delta_i)$ donde

$$\tilde{T} = \text{mín}\{T, Y\}, \qquad \delta_i = \begin{cases} 1, & \text{si } T \leq Y \\ 0, & \text{si } T > Y \end{cases}$$

donde $\tilde{T}$ es el tiempo observado y Δ_i es el indicador de censura.

Por la ecuación (2.15), la función de verosimilitud para un conjunto de datos en estudio y con censura por la derecha, viene dado de esta forma:

$$L(\theta) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{3.12}$$

3.3. Estimación e Inferencia del modelo

La función de log verosimilitud está dada por:

$$\begin{aligned} l(\theta) = & \sum_{i=1}^{n} [\delta_i log(p_{0i}(1-p_{0i}) \frac{\lambda_1 t_i^{\lambda_1 - 1}}{\lambda_2^{\lambda_1}} e^{-(\frac{t_i}{\lambda_2})^{\lambda_1}} \left\{ 1-(1-p_{0i}) e^{-(\frac{t_i}{\lambda_2})^{\lambda_1}} \right\}^{-2}) \\ & + (1-\delta_i) log(\frac{p_{0i}}{1-(1-p_{0i}) e^{-(\frac{t_i}{\lambda_2})^{\lambda_1}}})] \end{aligned}$$

Para el modelo sin covariables, el vector de parámetros es $\theta = (p_0, \lambda_1, \lambda_2)$
Las funciones de score estan dadas por

$$\frac{\partial l(\theta)}{\partial p_0} = \frac{n}{p_0} - \frac{1}{1-p_0} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} (\delta_i + 1) \frac{e^{-(\frac{t_i}{\lambda_2})^{\lambda_1}}}{1-(1-p_0) e^{-(\frac{t_i}{\lambda_2})\lambda_1}}$$

$$\frac{\partial l(\theta)}{\partial \lambda_1} = \sum_{i=1}^{n} \delta_i \left[\frac{1}{\lambda_1} + \left[1 - \left(\frac{t_i}{\lambda_2} \right)^{\lambda_1} \right] log \left(\frac{t_i}{\lambda_2} \right) \right] - \sum_{i=1}^{n} (\delta_i + 1) \frac{(1-p_0) log \left(\frac{t_i}{\lambda_2} \right) \left(\frac{t_i}{\lambda_2} \right)^{\lambda_1} e^{-\left(\frac{t_i}{\lambda_2} \right)^{\lambda_1}}}{1-(1-p_0) e^{-\left(\frac{t_i}{\lambda_2} \right)^{\lambda_1}}}$$

$$\frac{\partial l(\theta)}{\partial \lambda_2} = -\frac{\lambda_1}{\lambda_2} \sum_{i=1}^{n} \delta_i \left[1 - \left(\frac{t_i}{\lambda_2} \right)^{\lambda_1} \right] + \sum_{i=1}^{n} (\delta_i + 1) \frac{(1-p_0) t_i^{\lambda_1} \lambda_1 \lambda_2^{-\lambda_1 - 1} e^{-\left(\frac{t_i}{\lambda_2} \right)^{\lambda_1}}}{1-(1-p_0) e^{-\left(\frac{t_i}{\lambda_2} \right)^{\lambda_1}}}$$

Las segundas derivadas están dadas por:

$$\frac{\partial^2\ell(\theta)}{\partial p_0^2} = -\frac{n}{p_0^2} - \frac{1}{(1-p_0)^2}\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}(\delta_i+1)\frac{e^{-2\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}{\left(1-(1-p_0)e\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\right)^2}$$

$$\frac{\partial^2\ell(\theta)}{\partial \lambda_1^2} = -\sum_{i=1}^{n}\delta_i\left[\frac{1}{\lambda_1^2} + log^2\left(\frac{t_i}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\right] + \sum_{i=1}^{n}(\delta_i+1)\frac{(1-p_0)log^2\left(\frac{t_i}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\left[\left(\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}-1\right)e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+1-p_0\right]}{\left(e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+p_0-1\right)^2}$$

$$\frac{\partial^2\ell(\theta)}{\partial \lambda_2^2} = \frac{\lambda_1}{\lambda_2^2}\sum_{i=1}^{n}\delta_i - \sum_{i=1}^{n}\delta_i\lambda_1\lambda_2^{-2}(\lambda_1+1)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} + \sum_{i=1}^{n}(\delta_i+1)(1-p_0)\frac{\frac{\lambda_1}{\lambda_2^2}\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\left[\left(\lambda_1\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}-\lambda_1-1\right)e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+(1-p_0)\lambda_1+1-p_0\right]}{\left(e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+p_0-1\right)^2}$$

$$\frac{\partial\ell(\theta)}{\partial p_0\lambda_1} = \sum_{i=1}^{n}(\delta_i+1)\frac{log\left(\frac{t_i}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}{\left(e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+p_0-1\right)^2}$$

$$\frac{\partial\ell(\theta)}{\partial p_0\lambda_2} = -\sum_{i=1}^{n}(\delta_i+1)\frac{\left(\frac{\lambda_1}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}{\left(e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+p_0-1\right)^2}$$

$$\frac{\partial\ell(\theta)}{\partial \lambda_1\lambda_2} = \frac{1}{\lambda_2}\sum_{i=1}^{n}\delta_i\left[\left(\lambda_1 log\left(\frac{t_i}{\lambda_2}\right)+1\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}-1\right] + \sum_{i=1}^{n}(\delta_i+1) \cdot \frac{(p_0-1)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\left[\left(\lambda_1 log\left(\frac{t_i}{\lambda_2}\right)\left(\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}-1\right)-1\right)e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+(1-p_0)\left(\lambda_1 log\left(\frac{t_i}{\lambda_2}\right)-1\right)\right]}{\lambda_2\left(e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+p_0-1\right)^2}$$

En el caso con covariables, el vector de parámetros esta dado por $\theta = (\beta_0, \beta_1, ..., \beta_k, \lambda_1, \lambda_2)$

Las funciones de score para los parámetros están dados por:

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \lambda_1} &= \sum_{i=1}^{n}\delta_i\left[\frac{1}{\lambda_1}+\left[1-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\right]log\left(\frac{t_i}{\lambda_2}\right)\right]-\sum_{i=1}^{n}(\delta_i+1)\frac{log\left(\frac{t_i}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}{1+e^{x_i^T\beta}-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}\\
\frac{\partial l(\theta)}{\partial \lambda_2} &= -\frac{\lambda_1}{\lambda_2}\sum_{i=1}^{n}\delta_i\left[1-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\right]+\sum_{i=1}^{n}(\delta_i+1)\frac{(\frac{\lambda_1}{\lambda_2})(\frac{t_i}{\lambda_2})^{\lambda_1}e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}{1+e^{x_i^T\beta}-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}\\
\frac{\partial l(\theta)}{\partial \beta_0} &= -\sum_{i=1}^{n}(\delta_i+1)\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}+1\\
\frac{\partial l(\theta)}{\partial \beta_k} &= -\sum_{i=1}^{n}(\delta_i+1)\frac{e^{x_i^T\beta}x_{ik}}{1+e^{x_i^T\beta}-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}+x_{ik}
\end{aligned}$$

Las estimaciones de máxima verosimilitud (MLEs) $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k, \widehat{\lambda}_1, \widehat{\lambda}_2$ de los parámetros $\beta_0, \beta_1, ..., \beta_k, \lambda_1, \lambda_2$, se obtienen de las derivadas de la log-verosimilitud en (3.19) e igualando a cero:

$$\tfrac{\partial l(\theta)}{\partial \beta_0}=0, \tfrac{\partial l(\theta)}{\partial \beta_1}=0, ..., \tfrac{\partial l(\theta)}{\partial \beta_k}=0, \tfrac{\partial l(\theta)}{\partial \lambda_1}=0, \tfrac{\partial l(\theta)}{\partial \lambda_2}=0$$

Las segundas derivadas están dadas por:

$$\begin{aligned}
\frac{\partial^2 \ell(\theta)}{\partial \beta_0^2} &= -\sum_{i=1}^{n}(\delta_i+1)\frac{e^{x_i^T\beta}\left(1-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}\right)}{\left(1+e^{x_i^T\beta}-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}\right)^2}\\
\frac{\partial^2 \ell(\theta)}{\partial \beta_k^2} &= -\sum_{i=1}^{n}(\delta_i+1)\frac{e^{x_i^T\beta}x_{ik}^2\left(1-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}\right)}{\left(1+e^{x_i^T\beta}-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}\right)^2}\\
\frac{\partial^2 \ell(\theta)}{\partial \lambda_1^2} &= -\sum_{i=1}^{n}\delta_i\left[\frac{1}{\lambda_1^2}+log^2\left(\frac{t_i}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\right]\\
&\quad+\sum_{i=1}^{n}(\delta_i+1)\frac{log^2\left(\frac{t_i}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\left[(e^{x_i^T\beta}+1)\left(\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}-1\right)e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+1\right]}{\left(e^{x_i^T\beta+\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}-1\right)^2}\\
\frac{\partial^2 \ell(\theta)}{\partial \lambda_2^2} &= -\frac{\lambda_1}{\lambda_2^2}\sum_{i=1}^{n}\delta_i\left[(\lambda_1+1)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}-1\right]\\
&\quad+\sum_{i=1}^{n}(\delta_i+1)\frac{\left(\frac{\lambda_1}{\lambda_2^2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}\left[\left(e^{x_i^T\beta}+1\right)\left(\lambda_1\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}-\lambda_1-1\right)e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+\lambda_1+1\right]}{\left(e^{x_i^T\beta+\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}+e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}-1\right)^2}
\end{aligned}$$

$$\frac{\partial \ell(\theta)}{\partial \beta_0 \beta_k} = -\sum_{i=1}^{n} (\delta_i + 1) \frac{e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} \left(e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right) x_{ik}}{\left(e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} + e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right)^2}$$

$$\frac{\partial \ell(\theta)}{\partial \beta_k \beta_j} = -\sum_{i=1}^{n} (\delta_i + 1) \frac{e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} \left(e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right) x_{ik} x_{ij}}{\left(e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} + e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right)^2}$$

$$\frac{\partial \ell(\theta)}{\partial \lambda_1 \lambda_2} = \frac{1}{\lambda_2} \sum_{i=1}^{n} \delta_i \left[\left(\lambda_1 log\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} + 1\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} - 1\right]$$

$$-\sum_{i=1}^{n} (\delta_i + 1) \frac{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} \left[\left(e^{x_i^T\beta} + 1\right)\left(\lambda_1 log\left(\frac{t_i}{\lambda_2}\right)\left(\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} - 1\right) - 1\right) e^{\left(\frac{t_1}{\lambda_2}\right)^{\lambda_1}} + \lambda_1 log\left(\frac{t_i}{\lambda_2}\right) + 1\right]}{\lambda_2\left(e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} + e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right)^2}$$

$$\frac{\partial \ell(\theta)}{\partial \beta_0 \lambda_1} = \sum_{i=1}^{n} (\delta_i + 1) \frac{log\left(\frac{t_i}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}{\left(e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} + e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right)^2}$$

$$\frac{\partial \ell(\theta)}{\partial \beta_0 \lambda_2} = -\sum_{i=1}^{n} (\delta_i + 1) \frac{\left(\frac{\lambda_1}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}{\left(e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} + e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right)^2}$$

$$\frac{\partial \ell(\theta)}{\partial \beta_k \lambda_1} = \sum_{i=1}^{n} (\delta_i + 1) \frac{log\left(\frac{t_i}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} x_{ik}}{\left(e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} + e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right)^2}$$

$$\frac{\partial \ell(\theta)}{\partial \beta_k \lambda_2} = -\sum_{i=1}^{n} (\delta_i + 1) \frac{\left(\frac{\lambda_1}{\lambda_2}\right)\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} x_{ik}}{\left(e^{x_i^T\beta + \left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} + e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} - 1\right)^2}$$

Los intervalos de confianza de los estimadores de máxima verosimilitud se obtiene usando la distribución asintotica que establece

$$\sqrt{n}(\hat{\theta} - \theta) \sim N(0, I^{-1})$$

donde I es la matriz de información. En este caso la matriz de información no tiene forma explicita, por lo tanto la matriz observado (inversa de la matriz hessiana) es una buena aproximación.

Debido a que no es posible encontrar una solución analítica para las estimaciones de máxima verosimilitud, utilizaremos procedimiento de optimización numérica para encontrar $\widehat{\theta}$.

Capítulo 4

Estudio de Simulación

En este capítulo se desarrolla un estudio de simulación para evaluar el desempeño de las estimaciones de máxima verosimilitud de los parámetros del modelo. Esto se lleva a cabo calculando el sesgo, el error cuadrático medio y la cobertura de los intervalos de confianza al 95 %

4.1. Criterios para evaluar la simulación

Un criterio para evaluar el desempeño de los parámetros es el sesgo, el cual está definido por:

$$Sesgo = \frac{1}{M}\sum_{j=1}^{M}(\hat{\theta}_j - \theta) \tag{4.1}$$

donde $\hat{\theta}_j$ representa el estimado del parámetro de la $j-$ésima simulación y M es el número de simulaciones.
Adicionalmente, para evaluar el desempeño de los intervalos de confianza (IC) utilizamos el criterio de cobertura, definido por:

$$Cobertura = \frac{1}{M}\sum_{j=1}^{M} I(\theta \in [LI_j, LS_j]) \tag{4.2}$$

donde LI_j LS_j representan el límite inferior y superior del intervalo de confianza al 95 % generado con la j-ésima simulación.

4.2. Simulación sin covariables

El vector de parámetros del modelo de larga duración Weibull-Geométrica sin covariables es $\theta = (p_0, \lambda_1, \lambda_2)$. Para la simulación hemos considerado los siguientes valores de θ:

i) Escenario 1: $\theta = (0{,}5, 0{,}8, 1{,}1)$

ii) Escenario 2: $\theta = (0{,}8, 1{,}2, 0{,}6)$

Se considera tamaños de muestras $n = (250, 500, 1000)$ y llevamos a cabo $M = 1,000$ simulaciones en todos los escenarios considerados.

Para simular una muestra aleatoria bajo los supuestos del modelo seguimos el siguiente algoritmo:

i) Fijamos los valores de θ y n

ii) Para cada $i = 1, 2, ..., n$ se genera una variable aleatoria geométrica $M_i \sim Geo(p_0)$

iii) Para cada $i = 1, 2, ..., n$, si $M_i = 0$, entonces $T_i = \infty$ y si $M_i = m_i > 0$ entonces generar m_i variables aleatorias Weibull $W_1, W_2, ..., W_{m_i}$ con parámetros λ_1 y λ_2. El tiempo T_i esta definido por

$$T_i = \text{mín}\{W_1, W_2, ..., W_{m_i}\}$$

iv) Para cada $i = 1, 2, ..., n$, generar una variable uniforme $Y_i \sim U(0, 5)$.

v) Para cada $i = 1, 2, ..., n$, calcular $\tilde{T}_i = min(T_i, Y_i)$ y $\Delta_i = I(T_i \leq Y_i)$.

El cuadro 4.1 muestra los sesgos y las coberturas para intervalos de confianza del 95 % de las estimaciones obtenidas a través del estimador de máxima verosimilitud. En el escenario 1, la proporción de cura es de 0.5, la proporción de censura es 0.07 y la proporción de eventos de 0.42. En el escenario 2, la proporción de cura es de 0.8, la proporción de censura de 0.02 y la proporción de eventos de 0.18. Podemos observar que en ambos escenarios, los sesgos para cada parámetro se acercan a cero a medida que aumenta el tamaño de la muestra, cuando n es igual a 500 o 1,000 los sesgos parecen lo suficientemente pequeños y las coberturas estan alrededor del 95 %.

Cuadro 4.1: Sesgo y cobertura de las estimaciones de los parámetros considerando tamaños de muestra igual a 250, 500 y 1000

	$n = 250$		$n = 500$		$n = 1000$	
	Sesgo	Cobertura	Sesgo	Cobertura	Sesgo	Cobertura
Escenario 1						
$\lambda_1 = 0{,}8$	0,008	0.952	0,005	0.951	0,002	0.942
$\lambda_2 = 1{,}1$	0,135	0.896	0,064	0.934	0,018	0.933
$p_0 = 0{,}5$	-0,006	0.966	-0,003	0.953	-0,001	0.948
Escenario 2						
$\lambda_1 = 1{,}2$	0,041	0,952	0,012	0,938	0,011	0,950
$\lambda_2 = 0{,}6$	0,003	0,939	0,001	0,943	0,001	0,947
$p_0 = 0{,}8$	0,000	0,942	0,001	0,933	-0,001	0,948

4.3. Simulación con covariables

El vector de parámetros del modelo de larga duración Weibull Geométrica con covariables es $\theta = (\beta_0, \beta_1, \beta_2, \lambda_1, \lambda_2)$. En el proceso de simulación hemos considerado los siguientes valores de θ que definen cuatro escenarios:

i) Escenario 1: $\theta = (-1{,}2, 0{,}5, 1{,}1, 1{,}1, 1)$

ii) Escenario 2: $\theta = (-1{,}2, -0{,}5, 0{,}8, 1{,}2, 0{,}5)$

iii) Escenario 3: $\theta = (1{,}2, -0{,}5, 1{,}1, 1{,}2, 1)$

iv) Escenario 4: $\theta = (1, -1, -0{,}8, 0{,}9, 1{,}1)$

Se considera tamaños de muestras $n = (250, 500, 1000)$ y se llevaron a cabo $M = 1,000$ simulaciones. Entonces para cada simulación, considerando un tamaño de muestra n, se generan cada observación i-ésima $(i = 1, \ldots, n)$ de la siguiente manera:

i) Para cada $i = 1, \ldots, n$, generar $X_{1i} \sim Bin(n, 0{,}5)$, $X_{2i} \sim Bin(n, 0{,}4)$

iii) Para cada $i = 1, \ldots, n$ y dados los valores de $\beta_0, \beta_1, \beta_2$, generar

$$p_{0i} = \frac{e^{(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}}{1 + e^{(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}}$$

iv) Para cada $i = 1, 2, ..., n$, generar una variable aleatoria geométrica $M_i \sim Geo(p_0)$

v) Para cada $i = 1, \ldots, n$, si $M_i = 0$, entonces $T_i = \infty$. Si $M_i = m_i > 0$, generar $W_1, W_2, ..., W_{m_i}$ independientes y con distribución Weibull con parámetros λ_1 y λ_2. Entonces

$$T_i = \text{mín}\{W_1, W_2, ..., W_{m_i}\}$$

vi) Para cada $i = 1, \ldots, n$, generar una variable uniforme $Y_i \sim U(0, 5)$ que corresponde a los tiempos de censura

vii) Para cada $i = 1, \ldots, n$, calcular el tiempo obsevado $\tilde{T}_i = min(T_i, Y_i)$ y el indicador de censura $\Delta_i = I(T_i \leq Y_i)$.

En el cuadro 4.2 se presentan las simulaciones con covariables para los cuatro escenarios estudiados. Es importante mencionar que bajo estos escenarios se observan distintas proporciones de datos censurados (entre 0.07 y 0.03) y la proporción de eventos 0.55, 0.71, 0.18 y 0.39 para cada uno de los escenarios considerados. En cada escenario se puede observar que los sesgos de las estimaciones obtenidas se acercan a cero a medida que aumenta el tamaño de la muestra, asimismo las coberturas estan entre 92.2 % y 96.4 %.

El cuadro 4.3 presenta los resultados de las simulaciones incluyendo covariables para 3 escenarios. Los datos se generaron siguiendo los mismos pasos de la sección 4.2, con una variación en el paso 6 donde los tiempos de activación de los factores latentes se generan de una distribución log-normal. Esto se realizó para evaluar la robustez del modelo bajo una incorrecta especificación de éste. Del cuadro podemos observar que los sesgos de las estimaciones de los parámetros de la distribución log-normal son grandes para cualquier tamaño de muestra y las proporciones de cobertura en los distintos escenarios tienden a cero. En cuanto a los coeficientes de regresión, los sesgos se acercan a cero para los diferentes tamaños de muestra, solamente β_1 y β_2 presentan proporciones de cobertura mayor al 93 %. Por tanto, el rendimiento de las todas las estimaciones de los parámetros no es el adecuado en estos tres escenarios.

Cuadro 4.2: Sesgo y cobertura de las estimaciones de los parámetros que incluye covariables, considerando n igual a 250, 500 y 1000

	n=250		n=500		n=1000	
Escenario 1	Sesgo	Cobertura	Sesgo	Cobertura	Sesgo	Cobertura
$\lambda_1 = 1{,}1$	0,009	0,948	0,007	0,958	0,003	0,948
$\lambda_2 = 1$	0,031	0,939	0,021	0,940	0,001	0,938
$\beta_0 = -1{,}2$	-0,019	0,959	-0,020	0,938	0,000	0,938
$\beta_1 = 0{,}5$	0,009	0,946	0,004	0,945	0,004	0,941
$\beta_2 = 1{,}1$	0,020	0,939	0,013	0,954	-0,001	0,943
Escenario 2						
$\lambda_1 = 1{,}2$	0,012	0,954	0,006	0,941	0,002	0,955
$\lambda_2 = 0{,}5$	0,002	0,934	0,002	0,944	0,001	0,942
$\beta_0 = -1{,}2$	-0,003	0,953	-0,007	0,953	-0,002	0,950
$\beta_1 = -0{,}5$	-0,007	0,947	-0,002	0,942	0,000	0,943
$\beta_2 = 0{,}8$	0,002	0,948	0,002	0,948	0,006	0,940
Escenario 3						
$\lambda_1 = 1{,}2$	0,040	0,939	0,011	0,949	0,009	0,948
$\lambda_2 = 1$	0,021	0,928	0,005	0,945	-0,001	0,940
$\beta_0 = 1{,}2$	0,020	0,958	0,014	0,950	0,005	0,951
$\beta_1 = -0{,}5$	-0,021	0,961	-0,005	0,937	-0,002	0,954
$\beta_2 = 1{,}1$	0,022	0,956	0,027	0,951	0,008	0,967
Escenario 4						
$\lambda_1 = 0{,}9$	0,012	0,952	0,007	0,939	0,004	0,953
$\lambda_2 = 1{,}1$	0,104	0,922	0,053	0,936	0,018	0,931
$\beta_0 = 1$	-0,006	0,964	-0,003	0,944	-0,002	0,964
$\beta_1 = -1$	-0,016	0,949	-0,008	0,950	0,000	0,954
$\beta_2 = -0{,}8$	-0,005	0,954	-0,006	0,957	-0,006	0,962

Cuadro 4.3: Sesgo y cobertura bajo especificación incorrecta del modelo

	n=250		n=1000		n=500	
Escenario 1	Sesgo	Cobertura	Sesgo	Cobertura	Sesgo	Cobertura
$\lambda_1 = 1{,}1$	0,776	0,000	0,755	0,000	0,745	0,000
$\lambda_2 = 1$	1.471	0,029	1,401	0,000	1.406	0,000
$\beta_0 = -1{,}2$	0.347	0,704	0.369	0,522	0,360	0,276
$\beta_1 = 0{,}5$	0,009	0,947	0,009	0,941	0,009	0,947
$\beta_2 = 1{,}1$	0,048	0,950	0,018	0,953	0.025	0,944
Escenario 2						
$\lambda_1 = 1{,}2$	2,547	0,000	2.509	0,000	2.490	0,000
$\lambda_2 = 0{,}5$	2.598	0,000	2.593	0,000	2.589	0,000
$\beta_0 = -1{,}2$	0.240	0,844	0.244	0,763	0.254	0,575
$\beta_1 = -0{,}5$	0,000	0,948	-0,012	0,942	-0.012	0,938
$\beta_2 = 0{,}8$	0,041	0,949	0,030	0,946	0,026	0,937
Escenario 3						
$\lambda_1 = 0{,}9$	0,711	0,002	0,697	0,000	0,691	0,000
$\lambda_2 = 1{,}1$	1,136	0,125	1.068	0,002	1,036	0,000
$\beta_0 = 1$	0.333	0,830	0.320	0,712	0.321	0,461
$\beta_1 = -1$	-0,042	0,950	-0,013	0,946	-0.012	0,956
$\beta_2 = -0{,}8$	-0,035	0,940	-0,012	0,937	-0,007	0,953

Capítulo 5

Aplicación

El conjunto de datos de esta aplicación corresponde a clientes, de una entidad bancaria peruana, con tarjeta de credito. El evento de interés para este estudio es el tiempo a la cancelación de la tarjeta de credito adquirida y los factores asociado con la no cancelación. Los datos extraídos para el análisis corresponden a clientes que adquirieron y activaron una tarjeta de crédito de enero a diciembre del año 2015 (Cuadro 5.1). Se excluyeron todos los clientes con situación laboral independiente (ingresos no fijos) debido a que había un porcentaje alto de datos faltantes.

Cuadro 5.1: Cantidad de clientes a estudiar estratificados por los meses del año 2015

Meses	n^o de TC
Enero	1,397
Febrero	1,254
Marzo	1,027
Abril	781
Mayo	855
Junio	955
Julio	1,008
Agosto	1,329
Septiembre	1,417
Octubre	1,378
Noviembre	1,600
Diciembre	1,315
Total	**14,268**

En total se anlizaron a 14,268 clientes, a los cuales se les siguió por un periodo de 24 meses desde el momento que activaron su tarjeta de crédito hasta que se registró la cancelación de su tarjeta de crédito o su última observación. Del total de personas evaluadas, el 78.2 % fueron censurados (11,158), quiere decir que hay un importante porcentaje de clientes que no ha experimentado el evento al final del estudio. Para estudiar que factores podrian estar asociados con la probabilidad de ser inmune a la cancelación de la tarjeta de crédito, se evaluaron 16 variables mediante un modelo de regresión logístico.

La función **optim**, el algoritmo "L-BFGS-B", se eligió para la maximización, fue usada para maximizar la función de log verosimilitud.

5.1. Descripción de la muestra

En el cuadro 5.2 presentamos una breve descripción de las variables que serán incluidas en el modelo, las cuales se relacionarán con la proporción de clientes que no experimentan el evento de cancelación de tarjetas de crédito. Entre las variables consideradas para el modelo tenemos: ingreso del cliente en los últimos 3 meses, cantidad de productos que tiene en el banco (entre activos y pasivos), información de líneas de crédito otorgada y utilizada (en moneda nacional) en la entidad bancaria en estudio y en otros bancos del sistema financiero (últimos 3 meses), número de entidades donde el cliente ha adquirido tarjetas de crédito, información de la deuda total que tiene el cliente en el banco (incluye, productos como préstamos personales, créditos vehiculares e hipotecarios, entre otros), así como la deuda total que tiene el cliente en el sistema financiero. Variables de tenencia de productos en el banco, es decir si el cliente tiene o no cuenta de haberes, si tiene o no préstamos personales, si posee o no un crédito hipotecario y si cuenta o no con un crédito por convenio. Ratios de utilización de líneas de crédito y el ratio de cuota de cartera conocida también como share of wallet (sow) se han considerado también en el modelo.

Del cuadro 5.2 observamos que el ingreso mensual del cliente, la línea de crédito utilizada y la línea de crédito otorgada en los últimos 3 meses, presentan altas desviaciones, lo que indica la presencia de valores atípicos. Asimismo, respecto al ratio o porcentaje de linea de crédito utilizada en la entidad bancaria (cantidad de línea de crédito utilizada entre la cantidad de línea otorgada de la tarjeta de crédito), observamos que los clientes que no cancelan su tarjeta de crédito utilizan en promedio un mayor porcentaje de su línea de crédito (29 %) respecto a aquellos que si han cancelado su tarjeta (17 %).

Respecto a información de tenencia de productos del cliente, en promedio el cliente cuenta con 2 productos en el banco, no se observan diferencias en la cantidad de producto por cada grupo (clientes que cancelan y no cancelan tarjeta). Adicionalmente, respecto a clientes que tienen una cuenta de haberes y un crédito hipotecario, el grupo que no canceló tiene un mayor porcentaje en estos productos respecto al grupo de clientes que si cancelaron su tarjeta. En cuanto a clientes que tienen un préstamo personal y un crédito por convenio (préstamo cuyas cuotas mensuales son debitadas automáticamente de la cuenta de haberes del cliente), no se ven diferencias en el porcentaje de tenencia de estos productos.

Otra variable importante es el ratio de cuota de mercado conocida también como share of wallet (sow, este ratio indica que tanto el cliente invierte o gasta en productos o servicios del banco respecto a los gastos o inversiones que hace en total en el sistema financiero), se obtiene dividiendo la deuda que el cliente tiene en el banco entre la deuda total que tiene en el sistema financiero (todos los bancos con el que el cliente trabaja), según el cuadro observamos que el grupo de clientes que no cancelan su tarjeta de crédito invierten o gastan en promedio un 20 % en productos y servicios del banco, mientras que en el grupo que no cancelaron su tarjeta este porcentaje es menor (13 %).

Respecto a las líneas de crédito que el cliente tiene en otros bancos , se calculó también el ratio entre las líneas utilizadas y las líneas otorgadas, del cuadro se observa que el porcentaje de líneas utilizadas es similar en ambos grupos (promedio 29 %). Por otro lado, se tomó en cuenta también el número de entidades bancarias donde el cliente tiene tarjetas de crédito

(no incluye el banco que emitio la tarjeta), aquí observamos que el en grupo de clientes donde se ha dado el evento de interés, tienen tarjetas de crédito en 3 entidades bancarias en promedio, mientras que los clientes que no han realizado cancelaciones, trabajan con 2 entidades bancarias en promedio, donde tienen al menos una tarjeta de crédito.

Cuadro 5.2: Descripción de la muestra de clientes evaluados

Variables	Censurado 11,158 (78 %)	No Censurado 3,110 (22 %)	Total
Ingresos del cliente	4,255.64 (6,392.46)	3,704.29 (3,707.60)	4,135.46 (5,916.35)
Cantidad de productos	2 (1.31)	2 (1.19)	2 (1.29)
Linea de TC utilizada en el banco (*)	4,513.01 (10,336.52)	2,335.97 (6,312.91)	4,038.48 (9,646.02)
Linea de TC otorgada en el banco (*)	14,562.79 (18,094.44)	13,285.03 (14,444.85)	14,227.16 (17,219.49)
% Linea de TC utilizada en el banco (*)	0.29 (0.39)	0.17 (0.30)	0.26 (0.37)
Linea de TC utilizada en otros bancos	16970.45 (31,951.21)	8,766.35 (16,375.94)	15,182.2 (29,466.16)
Linea de TC otorgada en otros bancos	55,294.59 (64,082.90)	31,200.18 (39,243.05)	49,704.6 (60,120.20)
% Linea de TC utilizada (otros bcos)	0.29 (0.30)	0.28 (0.28)	0.29 (0.30)
Nro de entidades (cliente tiene TC)	2 (1.73)	3 (1.48)	2 (1.69)
Deuda total en el banco	16,949.83 (61,817.16)	7,839.54 (36,523.42)	14,964.06 (57,386.40)
Deuda total en el Sistema Financiero	87,795.45 (22,7594.60)	49,344.25 (15,0372.78)	79,414.23 (21,3745.97)
Ratio de cartera (Share of Wallet)	0.20 (0.32)	0.13 (0.27)	0.18 (0.31)
Cuenta de haberes			
No	9,957 (89.23 %)	2,868 (92.22 %)	12,825 (89.89 %)
Si	1,201 (10.76 %)	242 (7.78 %)	1,443 (10.11 %)
Préstamo personal			
No	10,110 (90.61 %)	2,851 (91.67 %)	12,961 (90.84 %)
Si	1,048 (9.39 %)	259 (8.33 %)	1,307 (9.16 %)
Crédito hipotecario			
No	10,315 (92.44 %)	2,980 (95.82 %)	13,295 (93.18 %)
Si	843 (7.56 %)	130 (4.18 %)	973 (6.82 %)
Crédito convenio			
No	10,791 (96.71 %)	2,995 (96.30 %)	13,786 (96.62 %)
Si	367 (3.29 %)	115 (3.70 %)	482 (3.38 %)

5.2. Modelo de larga duración Weibull-Geométrica sin covariables

La figura 5.1 muestra una comparación entre el estimador de Kaplan-Meier y el estimador asumiendo un modelo de larga duración Weibull-Geométrico para la función de supervivencia. Podemos observar que no hay mayor diferencia entre ambas estimaciones. Adicionalmente, el cuadro 5.3 muestra los estimadores de máxima verosimilitud de los parámetros, el error estándar y los intervalos de confianza al 95 % para p_0, λ_1 y λ_2. Asumiendo el modelo de larga duración Weibull-Geométrico, se estima que la proporción de clientes inmumes a cancelar su tarjeta es de 0.759 [IC95 %: 0.75-0.77].

5.3. Modelo de larga duración Weibull-Geométrica con covariables

Los resultados presentados en el cuadro 5.4 introducen los efectos de las covariables, través de la regresión logística, en la probabilidad de ser inmune a cancelar la tarjeta de crédito. Se muestran los estimadores de máxima verosimilitud de los parámetros, los errores estándar y sus respectivos intervalos de confianza al 95 %.

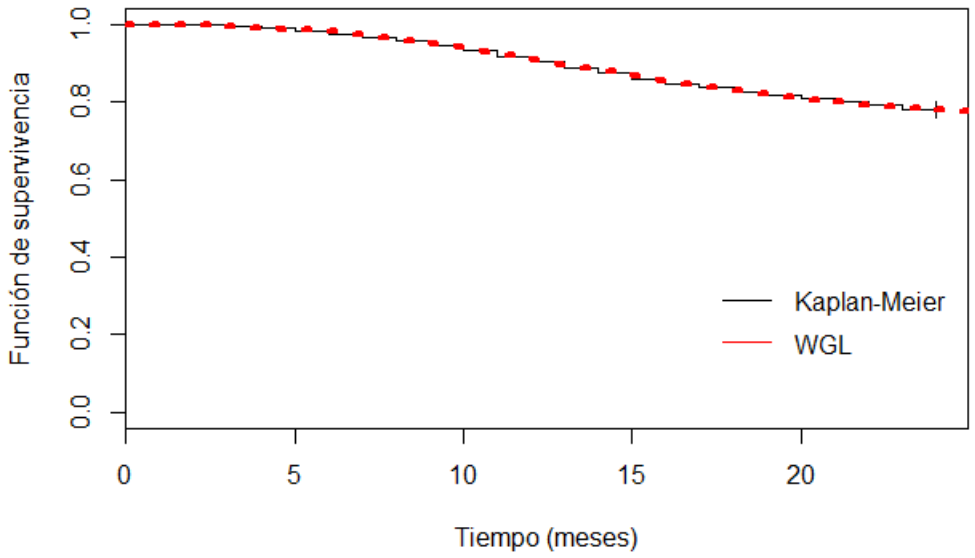

Figura 5.1: Función de supervivencia de Kaplan-Meier vs la supervivencia ajustada Weibull-Geométrico de larga duración (WGL)

Cuadro 5.3: Estimaciones sin covariables, errores estándar, intervalos de confianza al 95 % de los parámetros

Parámetros	Estimación	Error Est.	IC (95 %)
p_0	0.759	0.005	[0.750, 0.769]
λ_1	2.560	0.051	[2.460, 2.660]
λ_2	17.880	0.292	[17.308, 18.453]

Las covariables que se seleccionaron para el modelo de larga duración Weibull-Geométrico fueron: tenencia del producto cuenta de haberes (β_1), tenencia del producto hipotecario (β_2), porcentaje de línea de crédito utilizada por el cliente en el banco, promedio últimos 3 meses (β_3), porcentaje de línea de crédito utilizada por el cliente en otros bancos (no incluye al banco, promedio últimos 3 meses) (β_4), número de entidades bancarias donde el cliente tiene tarjetas de crédito (no incluye el banco en estudio)(β_5) y el ratio de cuota de cartera o share of wallet (β_6).

Respecto a las variables de tenencia de productos, de los resultados observamos que el odds de que un cliente que tiene una cuenta haberes no cancele su tarjeta de crédito es 1.29 [IC95 %: 1.11-1.49] veces mayor en comparación con un cliente que no posee cuenta de haberes, esto después de controlar por las otras variables. Este aumento en la proporción de clientes inmunes al evento de interés puede deberse que al tener una cuenta de haberes, el cliente tiene acceso a condiciones preferenciales con la tarjeta de crédito, como menor tasa de interés, promociones y descuentos en establecimientos de la preferencia del cliente.

En cuanto a clientes que si tienen un crédito hipotecario en el banco, el odds de que no cancelen su tarjeta de crédito aumenta en 56.8 % (OR=1.57, IC95 %: 1.27-1.93) en compa-

Cuadro 5.4: Modelo de regresión logistico para la probabilidad de ser susceptible al evento de interés

Parámetros	Estimación	Error Est.)	Odd Ratio	OR (95 % IC)
$\beta_{0(Intercepto)}$	1.242	0.047	3.461	]3.159, 3.793[
$\beta_{1(Hab:\ Si\ tiene)}$	0.252	0.075	1.287	]1.111, 1.491[
$\beta_{2(Hip:\ Si\ tiene)}$	0.450	0.106	1.568	]1.273, 1.932[
$\beta_{3(\%linea\ util.\ en\ el\ bco)}$	1.218	0.073	3.381	]2.929, 3.903[
$\beta_{4(\%linea\ util.\ otros\ bancos)}$	-0.164	0.076	0.849	]0.731, 0.986[
$\beta_{5(n°entidades\ TC\ sin\ el\ bco}$	-0.160	0.012	0.852	]0.832, 0.872[
$\beta_{6(Share\ of\ wallet)}$	0.184	0.086	1.203	]1.015, 1.425[
λ_1	2.582	0.051	2.582	]2.481, 2.683[
λ_2	18.135	0.308	18.135	]17.530, 18.739[

ración con aquellos clientes que no tienen hipotecario. Notemos que al ser este un producto que vincula al cliente por varios años (de 15 a 20 años) con el banco, podría darse el caso que por un tema de fidelización, el cliente haga mayor uso de los productos que tiene con el banco.

Respecto a la variable porcentaje de línea de crédito utilizada por el cliente en el banco, el odds de que el cliente no cancele su tarjeta es 3.4 (95IC %: 2.93-3.90) veces más por punto adicional en el porcentaje de utilización de la línea. Por tanto, esta variable es bastante influyente en la probabilidad de que el cliente no cancele su tarjeta de crédito, manteniendo fijas las otras variables del modelo.

Asimismo, para la variable correspondiente al ratio de cuota de cartera (share of wallet), del cuadro 5.6 observamos que el odds de que el cliente no cancele su tarjeta de crédito aumenta en 20 % (OR=1.2, 95 %IC: 1.02-1.42) por cada punto adicional en el share of wallet. Habíamos dicho anteriormente que este ratio nos indica cuanto invierte o gasta el cliente en productos o servicios del banco, por tanto, un incremento en este indicador, vincula o fideliza más cliente, haciendo que use más en este caso su tarjeta de crédito u otros productos que tenga aquí.

Por otro lado, respecto al porcentaje de líneas de crédito que utiliza el cliente en otros bancos, por cada punto adicional en el porcentaje de utilización de otras líneas de tarjetas, el chance de que el cliente no cancele su tarjeta disminuye en 15 % (OR=0.85, 95 %IC: 0.73-0.99).

Asimismo, de los resultados observamos que respecto al número de entidades bancarias donde el cliente tiene tarjetas de crédito (no incluye el banco que emitio la tarjeta de credito), el odds de que el cliente no cancele su tarjeta disminuye también en 15 % (OR=0.85, 95IC %: 0.83-0.87) por cada entidad adicional que el cliente tenga en el sistema financiero.

Las figuras 5.2 y 5.3 muestran las funciones de supervivencia para clientes que utilizan un porcentaje de su línea de crédito del 9.7 % y 53.5 % en el banco (cuantiles 60 y 75 %, respectivalmente), estratificados por la variable tenencia de crédito hipotecario, y manteniendo bajo control las variables porcentaje de líneas de crédito utilizadas en otros bancos (18.9 %, cuantil 50), número de entidades donde el cliente tiene tarjetas de crédito (2, cuantil 50), el share of wallet (18.5 %, media) y tenencia de cuenta de haberes. Para la figura 5.2 se consideró a clientes que tienen cuenta de haberes (Hab = Si), mientras que para la figura 5.3 a clientes

que no poseen cuenta de haberes (Hab = No).

De las figuras, hay que tener en cuenta que la función de supervivencia siempre es mayor cuando el cliente tiene un crédito hipotecario. Se observa que la probabilidad de supervivencia aumenta con un mayor valor en el porcentaje de línea de crédito utilizada en el banco, como era de esperar. Se ve también que cuando el cliente no posee cuenta de haberes ni crédito hipotecario, la probabilidad de no cancelar su tarjeta de crédito disminuye.

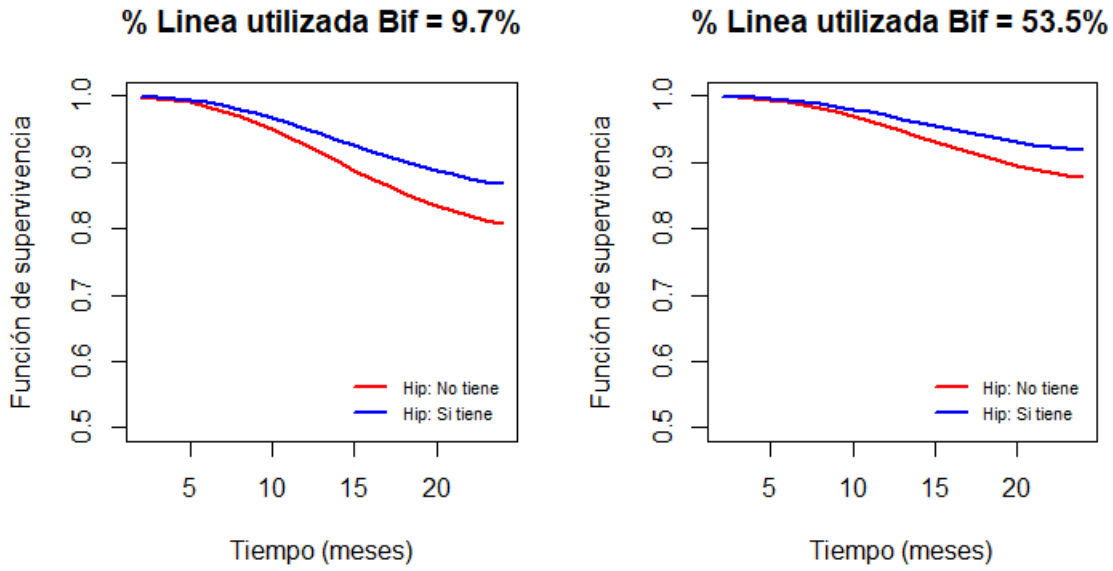

Figura 5.2: Función de supervivencia por tenencia de crédito hipotecario para clientes que utilizan un porcentaje de su línea del 9.7 y 53.5 %, tienen cuenta de haberes Hab=Si, línea utilizada en otros bancos 18.5 %, trabajan con 2 entidades donde tienen TC y tienen un share of wallet del 18.5 %

Las figuras 5.4 y 5.5 muestran también las funciones de supervivencia para clientes que utilizan un porcentaje de su línea de crédito del 9.7 % y 53.5 %, estratificados por el número de entidades donde el cliente tiene tarjetas de crédito, manteniendo bajo control las variables porcentaje de líneas de crédito utilizadas en otros bancos (18.9 %, cuantil 50), el share of wallet (18.5 %, media) y tenencia de productos, para la figura 5.4 se consideró a clientes que no tienen ambos productos (Hab = No, Hip = No), y en la figura 5.5 a clientes que sí cuentan con ambos productos (Hab = Sí, Hip=Sí).

Aquí la función de supervivencia siempre es mayor cuando el cliente solo trabaja con una entidad donde tiene al menos una tarjeta de crédito. En ambos gráficos se observa que la probabilidad de supervivencia aumenta con un mayor valor en el porcentaje de línea de crédito utilizada en el banco. Se ve también que cuando el cliente no posee cuenta de haberes ni crédito hipotecario y el porcentaje de línea de crédito que utiliza es baja (9.7 %), la probabilidad que no cancele su tarjeta disminuye.

5.4. Comparación con otros modelos

El cuadro 5.5 muestra una comparación entre los estimadores de máxima verosimilitud y los errores estándar de los parámetros, para los modelos de larga duración Weibull-Geométri-

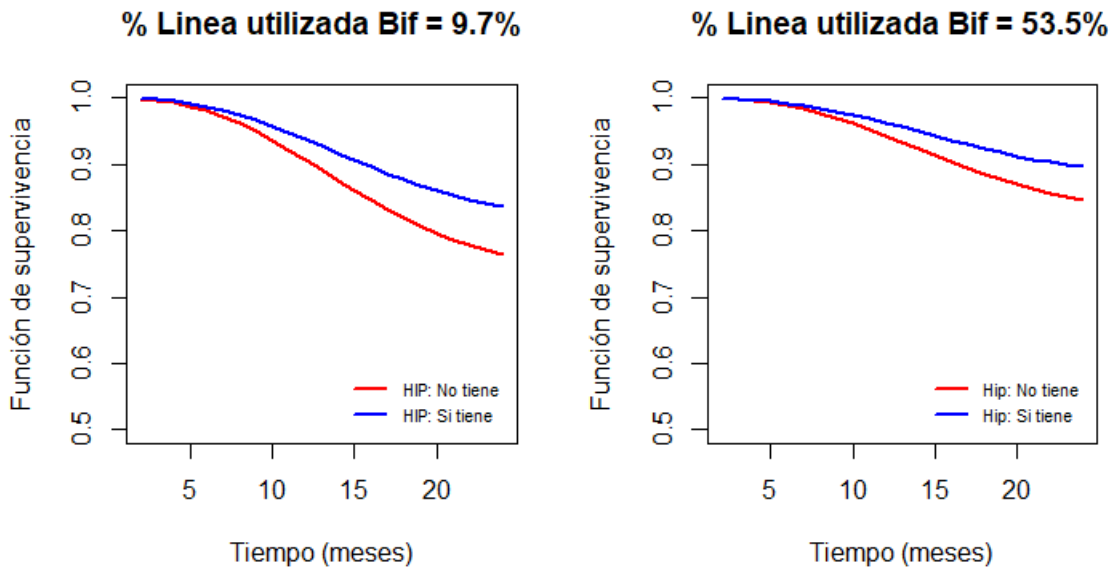

Figura 5.3: Función de supervivencia por tenencia de crédito hipotecario para clientes que utilizan un porcentaje de su linea del 9.7 y 53.5 %, tienen cuenta de haberes Hab=No, línea utilizada en otros bancos 18.5 %, trabajan con 2 entidades donde tienen TC y tienen un Sow del 18.5 %

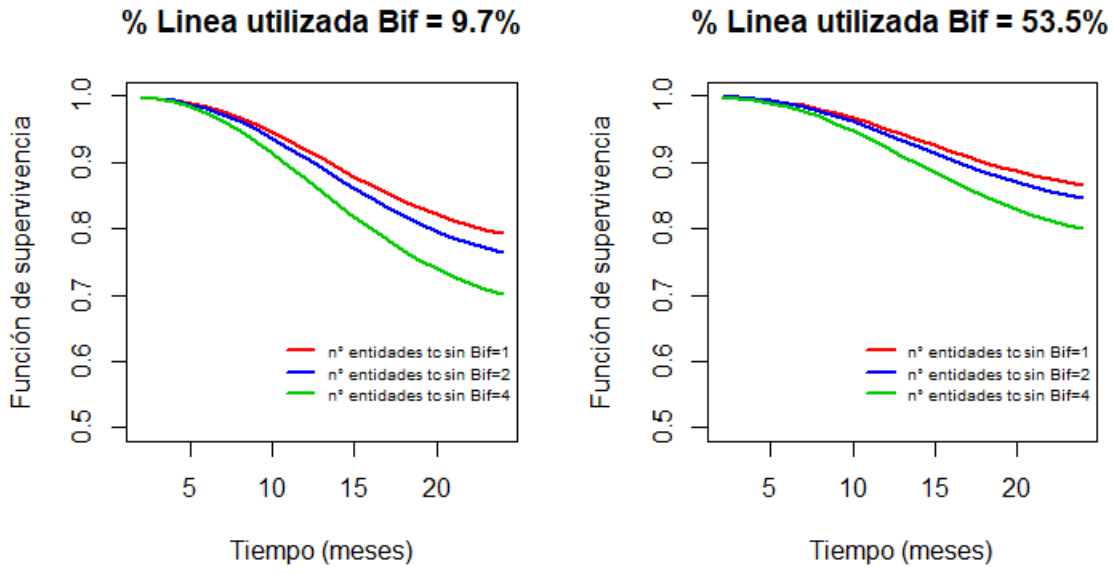

Figura 5.4: Función de supervivencia estratificado por el número de entidades donde el cliente tiene tarjeta de crédito y utiliza un porcentaje de su línea del 9.7 y 53.5 %, no tiene cuenta de haberes Hab=No, ni hipotecario Hip=No, línea utilizada en otros bancos del 18.5 % y tienen un sow del 18.5 %

co, Exponencial-Geométrico y Weibull. Estos dos últimas distribuciones son modelos particulares del modelo Weibull-Geométrico.

Las estimaciones de β_1 (alrededor de 0.25), β_2 (alrededor de 0.47), β_3 (alrededor de 1.20),

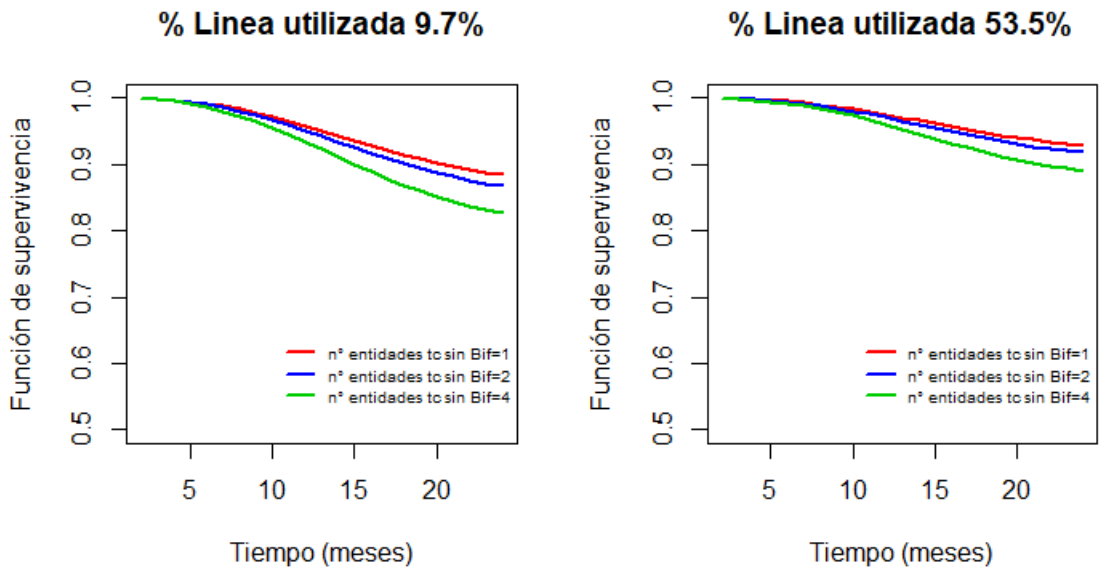

Figura 5.5: Función de supervivencia estratificado por el número de entidades donde el cliente tiene tarjeta de crédito y utiliza un porcentaje de su linea del 9.7 y 53.5 %, tiene cuenta de haberes Hab=Sí, y también hipotecario Hip=Sí, linea utilizada en otros bancos del 18.5 % y tienen un sow del 18.5 %

β_4 (alrededor de -0.19), β_5 (alrededor de -0.16) y β_6 (alrededor 0.16), están de acuerdo en los 3 modelos. Solo la estimación del beta0 sale negativo para el modelo Exponencial-Geométrico. Por otro lado, mencionar que la variable ratio de cuota de mercado (sow), no sale significativa en los modelos Weibull y Exponencial-Geométrico, los intervalos de confianza del odds ratio incluyen al 1.

Cuadro 5.5: Componente para modelar la probabilidad de ser inmume a cancelar la tarjeta de crédito via tres modelos de larga duración diferentes: Weibull-Geométrico (WGL), Weibull (WL) y Exponencial-Geométrico (EGL)

	WGL		WL		EGL	
Parámetros	Estimación	EE	Estimación	EE	Estimación	EE
$\beta_{0(Intercepto)}$	1.242	0.047	1.325	0.046	-2.485	0.177
$\beta_{1(Hab:\ Si\ tiene)}$	0.252	0.075	0.276	0.078	0.232	0.074
$\beta_{2(Hip:\ Si\ tiene)}$	0.450	0.106	0.479	0.109	0.472	0.107
$\beta_{3(\%linea\ util.\ en\ el\ bco)}$	1.218	0.073	1.244	0.075	1.144	0.073
$\beta_{4(\%linea\ util.\ otros\ bancos)}$	-0.164	0.076	-0.228	0.080	-0.192	0.075
$\beta_{5(n°entidades\ TC\ sin\ el\ bco}$	-0.160	0.012	-0.178	0.013	-0.143	0.012
$\beta_{6(Share\ of\ wallet)}$	0.184	0.086	0.138	0.089	0.165	0.086
λ_1	2.582	0.051	2.460	0.048		
λ_2	18.135	0.308	16.548	0.217	1190.646	205.973
EE: Error éstandar						

Para evaluar el desempeño de los modelos se ha usado el criterio de información Akaike (AIC), el criterio de Akaike Corregido (AICc) y el Criterio de Información Bayesiano (BIC). Del cuadro 5.6 se puede observar que el AIC, AICc y BIC más bajo está dado por el modelo Weibull-Geométrico, por lo tanto, concluimos que este modelo proporciona un mejor ajuste

de nuestro conjunto de datos.

Cuadro 5.6: Comparación de los modelos de larga duración Weibull-Geométrico (WGL), Weibull (WL) y Exponencial-Geométrico (EGL)

Criterio	WGL	WL	EGL
AIC	33222.69	33237.87	34480.61
AICc	33222.70	33237.88	34480.62
BIC	33290.78	33305.96	34541.14

Capítulo 6

Conclusiones

6.1. Conclusiones

En este trabajo hemos presentado y desarrollado el modelo de supervivencia de larga duración Weibull-Geométrica. Este model es una forma particular de generar un modelo de larga duración, ya que la función de supervivencia de los individuos susceptibles al evento de interés corresponde a una composición de distribuciones. El modelo se construyó sobre la base del esquema de la primera activación. Asimismo, se construyó la función de verosimilitud para el modelo sin y con covariables y se caracterizo el estimador de máxima verosimilitud.

Mediante un estudio de simulación se corroboro que los estimadores de máxima verosimilitud del modelo de larga duración Weibull-Geometrico tienen un buen desempeño (medido en sesgo porcentual y cobertura) aun para muestras pequeñas. Asimismo se realizó simulaciones con especificaciones incorrectas del modelo para validar la robutez del mismo.

Para este trabajo de tesis, el modelo el modelo de larga duración Weibull-Geométrico se aplicaron a un conjunto de datos correspondientes a clientes que adquirieron y activaron una tarjeta de crédito y fueron observados por un periodo de 24 meses, el evento de interés para este estudio es la cancelación de la tarjeta de crédito del cliente. En el modelo se incluyó información de covariables que nos ayudaron a identificar variables influyentes en la probabilidad de ser suceptible al evento de interés.

Se realizó una comparación del modelo propuesto con el modelo tradicional Weibull de larga duración y el modelo Exponencial-Geométrica de larga duración en términos del estadístico AIC. De los resultados obtenido el modelo Weibull-Geométrico proporciona un mejor ajuste del conjunto de datos en comparación con un modelo Weibull de larga duración.

6.2. Sugerencias para investigaciones futuras

Se han propuesto otros enfoques para modelos de larga duración. Chen et al. (1999), propuesto algunos modelos bayesianos para estimar la proporción de individuos no susceptibles o fracción de curados. Sy and Taylor (2000) discutio técnicas de máxima verosimilitud para el modelo de fracción de cura que tiene una estructura de riesgos proporcionales de Cox. Algunas sugerencias para futuros trabajos de investigación son

- Extender el estudio de modelos de larga duración bajo otros mecanismos de activación de la variable latente como por ejemplo última activación y aleatorio.
- Desarrollar la estimación bayesiana del modelo propuesto, y bajo los otros mecanismos.
- Desarrollar un método de selección de variables para el modelo de larga duración
- Realizar un análisis de influencia con el fin de mostrar la robustez del modelo propuesta ante la presencia de observaciones influyentes.

Apéndice A

Resultados teóricos

A.1. Distribución del $i-$ésimo estadístico de orden

Sean $Z_1, Z_2, ..., Z_M$ variables aleatorias independientes con función de distribución $G(\cdot)$ donde $Z_{(1)} \leq ... \leq Z_{(R)} \leq ... \leq Z_{(M)}$ son las estadísticas de orden de Z_i. Entonces definimos la siguiente variable aleatoria:

$$Y_i = \begin{cases} 1 & si \quad T_i \leq t \\ 0 & si \quad T_i > t \end{cases}$$

Las observaciones sobre estas variables aleatorias son:

i) $Y_i, Y_2, ..., Y_n$ son independientes con la misma distribución Bernoulli(p), donde $p = P(T_i \leq t) = G(t)$

ii) $Y_i + Y_2 + ... + Y_n \sim Bin(n, p)$

Por definición la función de distribución de la $i-$ésima estadística de orden esta dada por:

$$\begin{aligned} F_{Z_{(i)}}(z) &= P(Z_{(i)} \leq z) \\ & P\left(Y_1 + Y_2 + ... + Y_n \geq i\right) \\ & \sum_{j=i}^{m} \left(\begin{array}{c} m \\ j \end{array} \right) [G(t)]^j \left[1 - G(t)\right]^{m-j} \end{aligned} \tag{A.1}$$

A.2. Cuantil de la distribución Weibull-Geométrica

Invirtiendo la función acumulada de probabilidad dada en **??**:

$$\begin{aligned}
q &= \frac{1-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}}{1-(1-p)\,e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}} \\
q-q\,(1-p)\,e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} &= 1-e^{-\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}} \\
qe^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}-q\,(1-p) &= e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}-1 \\
e^{\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}}(1-q) &= log\,(1-q\,(1-p)) \\
\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1}+log\,(1-q) &= log\,(1-q\,(1-p)) \\
\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} &= log\,(1-q\,(1-p))-log\,(1-q) \\
\left(\frac{t_i}{\lambda_2}\right)^{\lambda_1} &= log\left(\frac{1-q\,(1-p)}{1-q}\right) \\
t_q &= \lambda_2\left\{log\left(\frac{1-q\,(1-p)}{1-q}\right)\right\}^{1/\lambda_1} \qquad \text{(A.2)}
\end{aligned}$$

Apéndice B

Anexo 2

B.1. Código simulación modelo Weibull Geométrico sin convariables

```
#### Simulación del modelo Weibull Geométrico sin covariables ####

sim <-1000
n <-c(250,500,1000)

## Generar data

  gen.data <- function(n){
  p0      <- 0.75
  lambda1 <- 2.56
  lambda2 <- 17.88
  tiempo  <- numeric(n)
  M       <- numeric(n)
  for (j in 1:n)
  {
    M[j]<-rgeom(1,p0)
    tiempo[j]<-ifelse(M[j]==0,Inf,min(rweibull(M[j],lambda1,lambda2)))
  }
  tc <- runif(n,0,5)
  t  <- pmin(tiempo,tc)
  status <- as.numeric(tiempo <= tc)
  dat <- data.frame(M=M,tiempo=tiempo,tc=tc,t=t,status=status)
  dat
  }

## Función de log-verosimilitud

llf= function(param,dat=dat)
{
  p0      <- param[1]
  lambda1 <- param[2]
  lambda2 <- param[3]

  log.fWGL <- log(p0) + log(1-p0) + log(lambda1) + (lambda1-1)*log(dat$t)
```

```
              - lambda1*log(lambda2) - (dat$t/lambda2)^lambda1
              - 2*log(1-(1-p0)*exp(-(dat$t/lambda2)^lambda1))
  log.SWGL <- log(p0) -log(1-(1-p0)*exp(-(dat$t/lambda2)^lambda1))
  llValue  = -sum(dat$status*log.fWGL+(1-dat$status)*log.SWGL,na.rm=T)
  return(llValue)
}

## Simulaciones ##

inicial <- c(0.85,1.0,0.5)

for (h in 1:length(n)){
  for(i in 1:sim)
  {
    dat <- gen.data(n[h])
    P.cura[i,h]<-sum(dat$M==0)/n[h]
    tiempos.cens[i,h]<-sum(dat$status==0)/n[h]
    censura.real[i,h]<-tiempos.cens[i,h]-P.cura[i,h]
    P.eventos[i,h]=sum(dat$status==1)/n[h]

# Maximizamos la función log-verosimilitud

 res<- optim(par=inicial,fn=llf,dat=dat,hessian=TRUE)

    H <- res$hessian           ## Matriz hessiana
    IFO <- solve(H)            ## Matriz de Información de Fisher observada para el modelo WGL
    ee_WGL <- sqrt(diag(IFO))  ## Medidas de los errores estándar de estimación

# Intervalos de confianza de los parámetros

    alpha=0.05
    z<-qnorm(1-alpha/2)

    # IC para beta0
    li.p0<-res$par[1]-z*ee_WGL[1]
    ls.p0<-res$par[1]+z*ee_WGL[1]

    # IC para lambda1
    li.lambda1<-res$par[2]-z*ee_WGL[2]
    ls.lambda1<-res$par[2]+z*ee_WGL[2]

    # IC para lambda2
    li.lambda2<-res$par[3]-z*ee_WGL[3]
    ls.lambda2<-res$par[3]+z*ee_WGL[3]

    estimacionWGL[i,(3*h-2):(3*h)]<-res$par
    e_estandarWGL[i,(3*h-2):(3*h)]<- ee_WGL
    IntervalosWGL[i,(6*h-5):(6*h)]<-c(li.p0,ls.p0,li.lambda1,ls.lambda1,li.lambda2,ls.lambda2)
```

```
  }
}
```

B.2. Código simulación modelo Weibull Geométrico con convariables

```
#### Simulación del modelo Weibull Geométrico con covariables ####

library(boot)
sim <-1000
n <-c(250,500,1000)

## Generar data
  gen.data <- function(n){
  x1       <- rbinom(n,1,0.5)
  x2       <- rbinom(n,1,0.4)
  beta0    <- 1 ; beta1 <- -1;  beta2 <- -0.8
  p0       <-  inv.logit(beta0 + beta1*x1 + beta2*x2)
  lambda1  <- 0.9
  lambda2  <- 1.1
  tiempo   <- numeric(n)
  M        <- numeric(n)
  for (j in 1:n)
  {
    M[j]=rgeom(1,p0[j])
    tiempo[j]=ifelse(M[j]==0,Inf,min(rweibull(M[j],lambda1,lambda2)))
  }
  tc <- runif(n,0,5)
  t  <- pmin(tiempo,tc)
  status <- as.numeric(tiempo <= tc)
  dat <- data.frame(M=M,tiempo=tiempo,tc=tc,t=t,status=status,x1=x1,x2=x2)
  dat
}

## Función de log-verosimilitud

llf= function(param,dat=dat)
{
  beta0   <- param[1]
  beta1   <- param[2]
  beta2   <- param[3]
  lambda1 <- param[4]
  lambda2 <- param[5]

  p0       <- inv.logit(beta0 + beta1*dat$x1 + beta2*dat$x2)
  log.fWGL <- log(p0) + log(1-p0) + log(lambda1) + (lambda1 - 1)*log(dat$t) -
              lambda1*log(lambda2) - (dat$t/lambda2)^lambda1 -
              2*log(1-(1-p0)*exp(-(dat$t/lambda2)^lambda1))
```

```
  log.SWGL <- log(p0) -log(1-(1-p0)*exp(-(dat$t/lambda2)^lambda1))
  llValue  <- -sum(dat$status*log.fWGL+(1-dat$status)*log.SWGL,na.rm=T)
  return(llValue)
}

## Simulaciones ##

inicial <- c(c(-0.5,-0.5,1),0.5,0.5)

for (h in 1:length(n)){
  for(i in 1:sim)
  {
  dat <- gen.data(n[h])
  P.cura[i,h]<-sum(dat$M==0)/n[h]
  tiempos.cens[i,h]<-sum(dat$status==0)/n[h]
  censura.real[i,h]<-tiempos.cens[i,h]-P.cura[i,h]
  P.eventos[i,h]<-sum(dat$status==1)/n[h]

# Maximizamos la función log-verosimilitud

  res<-optim(par=inicial,fn=llf,dat=dat,method="L-BFGS-B",lower=c(-Inf,-Inf,-Inf,-Inf,-Inf),
             upper=c(Inf,Inf,Inf,Inf,Inf),hessian=TRUE)

# Varianzas y errores estándar

  H   <- res$hessian        ## Matriz hessiana
  IFO <- solve(H)           ## Matriz de Información de Fisher observada para el modelo WGL
  ee_WGL <- sqrt(diag(IFO)) ## Medidas de los errores estándar de estimación

# Intervalos de confianza de los parámetros

  alpha<-0.05
  z<-qnorm(1-alpha/2)

  # IC para beta0
  li.beta0<-res$par[1]-z*ee_WGL[1]
  ls.beta0<-res$par[1]+z*ee_WGL[1]

  # IC para beta1
  li.beta1<-res$par[2]-z*ee_WGL[2]
  ls.beta1<-res$par[2]+z*ee_WGL[2]

  # IC para beta2
  li.beta2<-res$par[3]-z*ee_WGL[3]
  ls.beta2<-res$par[3]+z*ee_WGL[3]

  # IC para lambda1
  li.lambda1<-res$par[4]-z*ee_WGL[4]
```

```
  ls.lambda1<-res$par[4]+z*ee_WGL[4]

  # IC para lambda2
  li.lambda2<-res$par[5]-z*ee_WGL[5]
  ls.lambda2<-res$par[5]+z*ee_WGL[5]

  estimacionWGL[i,(5*h-4):(5*h)]<-res$par
  e_estandarWGL[i,(5*h-4):(5*h)]<- ee_WGL
  IntervalosWGL[i,(10*h-9):(10*h)]<-c(li.beta0,ls.beta0,li.beta1,ls.beta1,li.beta2,ls.beta2,
                                      li.lambda1,ls.lambda1,li.lambda2,ls.lambda2
  }
}
```

B.3. Código estimación modelo Weibull Geométrico con convariables

```
library(boot)
library(survival)

data_churn<-read.csv(file.choose(),sep=";")

## Función de log-verosimilitud

llf <- function(param,dat=data_churn)
{
  beta0   = param[1]
  beta1   = param[2]
  beta2   = param[3]
  beta3   = param[4]
  beta4   = param[5]
  beta5   = param[6]
  beta6   = param[7]
  lambda1 = param[8]
  lambda2 = param[9]

  p0       <- inv.logit(beta0 + beta1*dat$HAB + beta2*dat$HIP
                        + beta3*dat$Porc_Linea_Utilizada_TC_3m_banbif
                        + beta4*dat$Porc_Linea_Utilizada_TC_3m_sin_banbif
                        + beta5*dat$nro_entidades_TC_sin_banbif + beta6*dat$sow)
  log.fWGL <- log(p0) + log(1-p0) + log(lambda1) + (lambda1 - 1)*log(dat$tiempo_meses) -
              lambda1*log(lambda2) - (dat$tiempo_meses/lambda2)^lambda1 -
              2*log(1-(1-p0)*exp(-(dat$tiempo_meses/lambda2)^lambda1))
  log.SWGL <- log(p0) -log(1-(1-p0)*exp(-(dat$tiempo_meses/lambda2)^lambda1))
  llValue  = -sum(dat$status*log.fWGL+(1-dat$status)*log.SWGL,na.rm=T)
  return(llValue)
}

 inicial <- c(c(-0.5,-0.32,-0.58,-0.97,-0.10,0.11,-0.78),3.75,22.2)
```

```
# Maximizamos la función log-verosimilitud

 res<-optim(par=inicial,fn=llf,dat=data_churn,method="L-BFGS-B",lower=c(-Inf,-Inf,-Inf,-Inf,
     -Inf,-Inf,-Inf,-Inf,-Inf), upper=c(Inf,Inf,Inf,Inf,Inf,Inf,Inf,Inf,Inf),hessian=TRUE)

# Varianzas y errores estándar

 parametros<-res$par
 H   <- res$hessian        ## Matriz hessiana
 IFO <- solve(H)           ## Matriz de Información de Fisher observada para el modelo WGL
 ee_WGL <- sqrt(diag(IFO)) ## Medidas de los errores estándar de estimación

 AIC  <- -2*(-res$value)+2*length(parametros)
 BIC  <- -2*(-res$value)+log(n)*length(parametros)
 AICc <- AIC + 2*length(parametros)*(length(parametros)+1)/(n-length(parametros)-1)
```

Bibliografía

Adamidis, K., Dimitrakopoulou, T. and Loukas, S. (2005). On an extension of the exponential-geometric distribution, *Statistics & probability letters* **73**(3): 259–269.

Adamidis, K. and Loukas, S. (1998). A lifetime distribution with decreasing failure rate, *Statistics & Probability Letters* **39**(1): 35–42.

Barreto-Souza, W., de Morais, A. L. and Cordeiro, G. M. (2011). The weibull-geometric distribution, *Journal of Statistical Computation and Simulation* **81**(5): 645–657.

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

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

Cancho, V. G., Louzada-Neto, F. and Barriga, G. D. (2011). The poisson-exponential lifetime distribution, *Computational Statistics & Data Analysis* **55**(1): 677–686.

Chen, M.-H., Ibrahim, J. G. and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association* **94**(447): 909–919.

Chen, M. H., Shao, Q. M. and Ibrahim, J. G. (2001). *Monte Carlo Methods in Bayesian Computation*, Springer.

Cooner, F., Banerjee, S., Carlin, B. P. and Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes, *Journal of the American Statistical Association* **102**(478): 560–572.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*, Chapman and Hall.

Farewell, V. T. (1977). A model for a binary variable with time-censored observations, *Biometrika* **64**(1): pp. 43–46.
URL: *http://www.jstor.org/stable/2335768*

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics* pp. 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk?, *Canadian Journal of Statistics* **14**(3): 257–262.

Ghitany, M., Maller, R. and Zhou, S. (1995). Estimating the proportion of immunes in censored samples: a simulation study, *Statistics in medicine* **14**(1): 39–49.

Goldman, A. I. (1984). Survivorship analysis when cure is a possibility: a monte carlo study, *Statistics in Medicine* **3**(2): 153–163.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American statistical association* **53**(282): 457–481.

Kuş, C. (2007). A new lifetime distribution, *Computational Statistics & Data Analysis* **51**(9): 4497–4509.

Marshall, A. W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and weibull families, *Biometrika* **84**(3): 641–652.

Sposto, R. (2002). Cure model analysis in cancer: an application to data from the children's cancer group, *Statistics in medicine* **21**(2): 293–312.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model, *Biometrics* **56**(1): 227–236.

Tong, E. N., Mues, C. and Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default, *European Journal of Operational Research* **218**(1): 132–139.

Yakovlev, A. Y. (1996). Threshold models of tumor recurrence, *Mathematical and computer modelling* **23**(6): 153–164.

Yakovlev, A. Y., Asselain, B., Bardou, V., Fourquet, A., Hoang, T., Rochefediere, A. and Tsodikov, A. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer, *Biometrie et analyse de donnees spatio-temporelles* **12**: 66–82.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic models of tumor latency and their biostatistical applications*, World Scientific.

Yin, G. and Ibrahim, J. G. (2005). Cure rate models: a unified approach, *Canadian Journal of Statistics* **33**(4): 559–570.