PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

El modelo de larga duración Exponencial-Poisson

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:
Julia Elena Gonzales Rodriguez

Asesor: Victor Giancarlo Sal y Rosas Celi

Miembros del jurado:
Dr. Luis Valdivieso Serrano
Dr. Victor Giancarlo Sal y Rosas Celi
Dra. Rocío Maehara Aliaga

Lima, 12 de noviembre de 2018

Dedicatoria

A Julia, mi mamita, que siempre gua mis pasos. A mis padres por darme siempre su apoyo incondicional.

Agradecimentos

En primer lugar quiero agradecer a mi asesor Giancarlo Sal y Rosas por la orientación que me ha dado durante este proyecto porque sin sus consejos y enseñanzas esto no hubiera sido posible. Muchas gracias por toda su paciencia y por motivarme a ser constante en este camino. De igual manera a mi profesor José Julio Flores por guiarme en los inicios de este proyecto y brindarme todo su apoyo.

También quiero agradecer de forma muy especial a mi mamá y papá por su apoyo incondicional que me han dado en esta etapa. Gracias por acompañar cada uno de mis pasos y hacer que este camino sea menos complicado. Agradezco a mi familia por siempre estar pendiente de mi y alentarme en todo momento. Gracias a Elvis, mi gran compañero, por siempre darme la mano y apoyarme en todo lo que ha estado a su alcance.

Muchas gracias a mis profesores y amigos de la maestra por todo lo que he aprendido de ustedes en estos dos años de clases.

Finalmente gracias a mis amigos del trabajo, gracias a ustedes conocí el camino a la Estadística del cual no pienso alejarme.

Resumen

En esta tesis se introducir y estudiar el modelo de supervivencia de larga duración Exponencial-Poisson. Este modelo permite estudiar el tiempo hasta la ocurrencia de un evento de inters cuando se asume que existe una fracción de unidades de la población inmunes a la ocurrencia de este evento. El modelo descrito en esta tesis es un modelo de mixtura que usa la distribución Exponencial-Poisson para modelar el tiempo a la ocurrencia del evento de inters en la sub población suceptible al evento de interés. Además se plantea un modelo de regresión logística sobre la probabilidad de ser inmune al evento de interés. Se realiza un estudio de simulación en el cual a través del sesgo porcentual y cobertura se comprobó la buena performancia del modelo. Finalmente, el modelo es aplicado sobre una muestra de clientes morosos de una entidad del sistema financiero Peruano donde el evento de interés es la cancelación de dicha deuda.

Palabras-clave: modelo larga duración, mixtura, exponencial-poisson, censura.

Abstract

In this thesis the long-term survival model Exponential-Poisson will be introduced and discussed. This model allows to study the time until the occurrence of an event of interest when it is assumed that there is a fraction of the population that is immune to the occurrence of this event. The studied model is a mixture model that assumes that the time to the event among susceptible follows a Exponential-Poisson distribution and that the probability of being inmune to the event of interes is explained by a set of covariates via a logistic regression model. A simulation study was carried out in which the good performance of the model was checked through the percentage bias and 95 % coverage. Finally, the model is applied to a sample of a Peruvian finantial entity where the event of interest is the cancellation of the debt.

Keywords: long-term survival model, mixture model, exponential-poisson, censor.

Índice general

Lista de abreviaturas VII

Lista de símbolos VIII

Índice de figuras IX

Índice de cuadros X

1. Introducción 1

2. Conceptos preliminares 3
- 2.1. Distribución Exponencial . . . 3
- 2.2. Distribución Poisson . . . 3
- 2.3. Función hipergeométrica generalizada . . . 4
- 2.4. Distribución Exponencial-Poisson . . . 4
- 2.5. Datos censurados . . . 7
- 2.6. Estimador de Kaplan - Meier . . . 7
- 2.7. Modelos de regresión paramétrica . . . 8
- 2.8. El modelo de mixtura . . . 8

3. El modelo de larga duración Exponencial-Poisson 10
- 3.1. Modelo . . . 10
- 3.2. Estructura de datos . . . 12
- 3.3. Estimación e inferencia . . . 13
- 3.4. Comparación de modelos . . . 15

4. Simulación 16
- 4.1. Consideraciones para la simulación . . . 16
- 4.2. Resultados . . . 17

5. Aplicación 19
- 5.1. Descripción de los clientes . . . 19
- 5.2. Modelo final . . . 21
- 5.3. Comparación de modelos . . . 23

6. Conclusiones **26**
6.1. Conclusiones 26
6.2. Investigaciones futuras 26

A. Resultados teóricos **27**

B. Implementacion del codigo en R del modelo **28**

C. Implementación del código en R de la simulación **30**

Bibliografía **36**

Lista de abreviaturas

df	Función de densidad de probabilidad .
cdf	Función de distribución acumulada.
va	Variable aleatoria

Lista de símbolos

μ	Media.
T	Tiempo hasta que ocurra un evento de interés.
P	Probabilidad.
f, g	Función de densidad.
F, G	Función acumulada.
S	Función de supervivencia.
p_0	Proporción de unidades inmunes al evento de interés.
M	Factor de riesgo latente.

Índice de figuras

2.1. Función de densidad del modelo Exponencial-Poisson (γ, $\lambda = 1$) 5
2.2. Función de supervivencia del modelo Exponencial-Poisson (γ, $\lambda = 1$) 5
2.3. Función de riesgo instantáneo del modelo Exponencial-Poisson (γ, $\lambda = 1$) 6
2.4. Modelo de mixtura 9

3.1. Densidad y función de supervivencia del modelo de larga duración Exponencial-Poisson 11
3.2. Función de riesgo instántaneo del modelo de larga duración Exponencial-Poisson . . 12

5.1. Función de supervivencia sin covariables: Kaplan-Meier vs. Exponencial-Poisson de larga duración 22
5.2. Función de supervivencia con una covariable: Ser casado 24

Índice de cuadros

4.1. Parámetros considerados en la simulación 16
4.2. Sesgo porcentual y cobertura de las estimaciones: modelo bien especificado . 17
4.3. Sesgo porcentual y cobertura bajo la incorrecta especificación del modelo . . 18

5.1. Características de la muestra (Parte A) 20
5.2. Características de la muestra (Parte B) 21
5.3. Componente de regresión logístico para el modelo de larga duración Exponencial - Poisson para evaluar factores asociados con el no pagar su deuda 23
5.4. Comparación de modelos de larga duración Exponencial-Poisson, larga duración Exponencial y Logístico 25

Capítulo 1

Introducción

En muchas ocasiones se desea estudiar el tiempo, denotado por T, hasta que ocurra un evento de interés a partir de ciertas variables que expliquen el hecho. Este es el problema que se aborda en el análisis de supervivencia (Hosmer *et al.*, 2011). Sin embargo, existen eventos que no ocurren en cierto grupo de las unidades de observación, esto es, $P(T = \infty) > 0$ y, por lo tanto, T es una variable aleatoria extendida (Ash, 1972, p. 209).

Tal como lo propone Boag (1949), el enfoque clásico para el estudio de datos de supervivencia de larga duración es el modelo de mixtura, en el cual se asume que la población de unidades de estudio está conformada por inmunes y susceptibles al evento de interés. En la presente tesis se discutira un modelo de larga duración donde la probabilidad de ser inmune será explicada por covariables mediantes una regresión logística y el tiempo a la ocurrencia del evento de interés se asumirá que sigue una distribución Exponencial-Poisson.

La distribución Exponencial-Poisson fue introducida por Kuş (2007). Posteriormente, diversos autores han realizado estudios de esta distribución. En particular, Karlis (2009) propone un algoritmo de Esperanza-Maximización anidado para la estimación de los parámetros de la distribución. Barreto-Souza & Cribari-Neto (2009) generaliza la distribución y presenta la función Exponencial-Poisson generalizada. Cancho *et al.* (2011) introduce la distribución Exponencial-Poisson con función de riesgo instántaneo creciente y Rodrigues *et al.* (2018) continua profundizando en dicha investigación. Finalmente, Louzada *et al.* (2012) introduce el concepto de cura en la distribución Exponencial-Poisson y presenta la distribución, inferencia, estudios de simulación y aplicación en una muestra de pacientes con cáncer de ovario.

El objetivo general de la tesis es estudiar, deducir y aplicar a un conjunto de datos reales el modelo de supervivencia de larga duración Exponencial-Poisson. De manera específica:

- Revisar conceptos preliminares del análisis de supervivencia para comprender el modelo propuesto.
- Presentar y deducir el modelo de larga duración Exponencial-Poisson asumiendo modelo de regresión logístico para la probabilidad de ser suceptible al evento de interés. Se discutirá el proceso de estimación mediante máxima verosimilitud.
- Aplicar el modelo de larga duración Exponencial-Poisson a una muestra de clientes de una entidad financiera peruana. Estos clientes dejaron de pagar dos meses en el pago

de su crédito adquirido en un banco. El objetivo es estimar el tiempo que transcurre hasta que este tipo de cliente cancela el crédito.

La presente tesis esta organizada de la siguiente forma. En el primer capítulo se pone en contexto el trabajo realizado, los objetivos y la organización del trabajo. En el segundo capítulo encontraremos algunos conceptos preliminares para poder comprender el modelo propuesto en el presente trabajo. Se presentará las distribuciones exponencial, poisson, exponencial-poisson y el modelo de mixtura. Adicionalmente, se describirá los modelos para datos censurados. El tercer capítulo está dedicado al modelo de larga duración Exponencial-Poisson. Se iniciará explicando los factores de riesgos latentes y los tiempos de activación para posteriormente ahondar en la derivación del modelo. En el cuarto capítulo se presenta un estudio de simulación con la finalidad de mostrar el desempeño del modelo propuesto considerando diferentes escenarios. En el quinto capítulo se presenta la aplicación del modelo a un conjunto de datos de una entidad del sistema financiero.

Capítulo 2

Conceptos preliminares

2.1. Distribución Exponencial

Sea T una variable aleatoria (v.a) no negativa. Se dice que T tiene una distribución exponencial, denotada por $T \sim Exp(\lambda)$, si su función de densidad de probabilidad está dada por

$$f(t) = \lambda e^{-\lambda t},\ t > 0 \tag{2.1}$$

donde $\lambda > 0$ es un parámetro del modelo.
Su función de distribución acumulada es:

$$F(t) = 1 - e^{-\lambda t},\ t > 0 \tag{2.2}$$

Su función de supervivencia esta dada por:

$$S(t) = 1 - F(t) = e^{-\lambda t},\ t > 0 \tag{2.3}$$

La media y la varianza de T están dadas por:

$$E(T) = \frac{1}{\lambda} \quad \text{y} \quad Var(T) = \frac{1}{\lambda^2}. \tag{2.4}$$

2.2. Distribución Poisson

Sea Y una variable aleatoria (v.a) no negativa. Se dice que Y tiene una distribución Poisson, denotada por $Y \sim P(\gamma)$, si su función de densidad de probabilidad está dada por

$$f(y) = e^{-\gamma}\,\frac{\gamma^y}{y!},\ y = 0, 1... \tag{2.5}$$

donde $\gamma > 0$ es un parámetro del modelo.
La media y la varianza de Y, están dadas por:

$$E(Y) = \gamma \quad \text{y} \quad Var(Y) = \gamma. \tag{2.6}$$

2.3. Función hipergeométrica generalizada

La función hipergeométrica es utilizada para expresar funciones características de probabilidad o momentos (Kilbas *et al.*, 2012). Esta función denotada por $F_{p,q}(\boldsymbol{a}, \boldsymbol{b}, \gamma)$ esta definida de la siguiente manera

$$F_{p,q}(\boldsymbol{a}, \boldsymbol{b}, \gamma) = \sum_{j=0}^{\infty} \frac{\prod_{i=1}^{p}(a_i)_j \gamma^j}{\prod_{i=1}^{q}(b_i)_j j!} \tag{2.7}$$

donde $\gamma > 0$, $p = 0, 1, \ldots$, $q = 0, 1, \ldots$. Los valores $(a)_j, (b)_j$ son símbolos de Pochhammer definidos como

$$(a)_j = \frac{\Gamma(a+k)}{\Gamma(a)} = a(a+1)\ldots(a+k-1)$$

donde Γ es la función gamma. En ese contexto, la función hipergeométrica generalizada se puede reformular como

$$F_{p,q}(\boldsymbol{a}, \boldsymbol{b}, \gamma) = \sum_{j=0}^{\infty} \frac{\gamma^j \prod_{i=1}^{p}\Gamma(a_i+j)\Gamma(a_i)^{-1}}{\Gamma(j+1)\prod_{i=1}^{q}\Gamma(b_i+j)\Gamma(b_i)^{-1}}$$

donde $\boldsymbol{a} = (a_1 \ldots, a_p)$ y $\boldsymbol{b} = (b_1 \ldots, b_q)$.

2.4. Distribución Exponencial-Poisson

Sea T una variable aleatoria (v.a) no negativa. Se dice que T tiene una distribución Exponencial-Poisson (Kuş, 2007), denotada por $T \sim EP(\gamma, \lambda)$, si su función de densidad de probabilidad está dada por

$$f(t) = \frac{\gamma\lambda e^{-\gamma-\lambda t+\gamma e^{-\lambda t}}}{1-e^{-\gamma}},\ t > 0 \tag{2.8}$$

donde $\gamma > 0$ y $\lambda > 0$ son parámetros del modelo.
En la Figura 2.1 se muestra la función de densidad del modelo para diferentes valores del parámetro γ. Es sencillo notar que cuando γ se aproxima a cero, la distribución Exponencial-Poisson converge a una distribución exponencial con parámetro λ:

$$\begin{aligned}
\lim_{\gamma\to 0} f(t) &= \lim_{\gamma\to 0} \frac{\gamma\lambda e^{-\gamma-\lambda t+\gamma e^{-\lambda t}}}{1-e^{-\gamma}} \\
&= \frac{\lambda(e^{-\gamma-\lambda t+\gamma e^{-\lambda t}} + \gamma e^{-\gamma-\lambda t+\gamma e^{-\lambda t}}(-1+e^{-\lambda t}))}{e^{-\gamma}} \\
&= \lambda e^{-\lambda t}
\end{aligned}$$

Su función de distribución acumulada está dada por:

$$F(t) = \frac{e^{\gamma e^{-\lambda t}} - e^{\gamma}}{1-e^{\gamma}},\ t > 0 \tag{2.9}$$

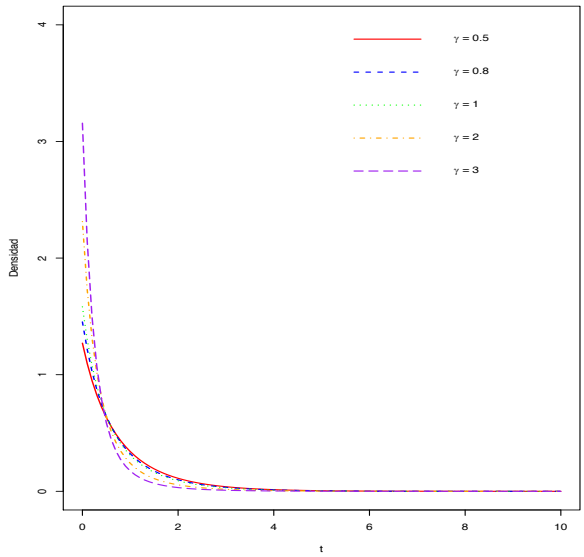

Figura 2.1: Función de densidad del modelo Exponencial-Poisson (γ, $\lambda = 1$)

Su función de supervivencia es:

$$S(t) = \frac{1 - e^{\gamma e^{-\lambda t}}}{1 - e^{\gamma}}, \ t > 0 \tag{2.10}$$

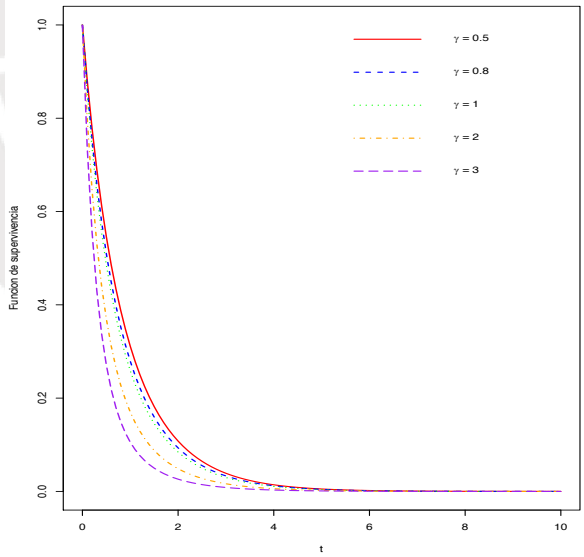

Figura 2.2: Función de supervivencia del modelo Exponencial-Poisson (γ, $\lambda = 1$)

De (2.1) y (2.2) se verifica que la función de riesgo instántaneo esta dada por:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\gamma\lambda e^{-\lambda+\lambda t+\gamma e^{-\lambda t}}(1 - e^{\gamma})}{(1 - e^{-\gamma})1 - e^{\gamma e^{-\lambda t}}}, \ t > 0 \tag{2.11}$$

La Figura 2.3 presenta el comportamiento de la función de riesgo instantaneo para diferentes

valores del parámetro γ. Se observa que conforme t aumenta, la función converge a el parámetro λ. Matemáticamente es sencillo mostrar que:

$$h(0) = \frac{\gamma\lambda e^1(1-e^{\gamma})}{(1-e^{-\gamma})(1-e^{\gamma}} = \frac{\lambda\gamma e^{\gamma}}{(e^{\gamma}-1)}$$

y

$$\lim_{t\to\infty} h(t) \simeq \lambda$$

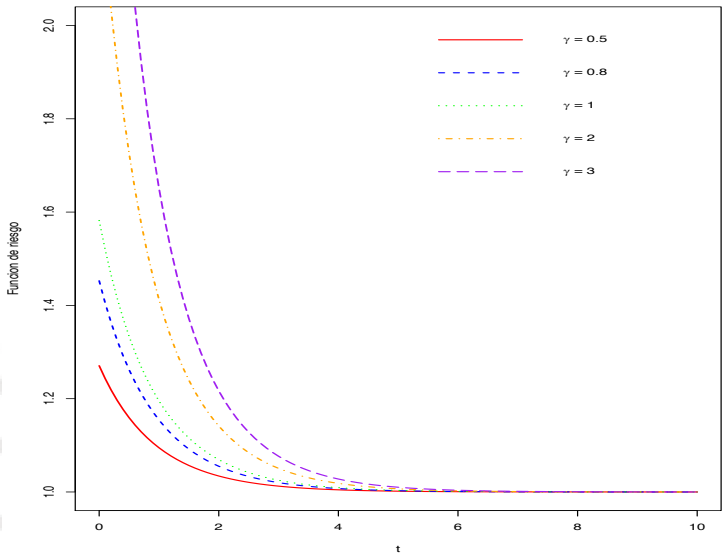

Figura 2.3: Función de riesgo instantáneo del modelo Exponencial-Poisson (γ, $\lambda = 1$)

Para construir la función de cuantiles debemos calcular la función inversa de la función acumulada (2.9). En particular, despejar la ecuación siguiente en función de t

$$F(t) = \frac{e^{\gamma e^{-\lambda t}} - e^{\gamma}}{1-e^{\gamma}} = z$$

resulta en

$$t = \frac{\log(\gamma) - \log(\log(z - e^{\gamma}(z-1)))}{\lambda}$$

y por ende la función inversa de la función de distrubución acumulada esta dada por

$$F^{-1}(t) = \frac{\log(\gamma) - \log(\log(t - e^{\gamma}(t-1)))}{\lambda} \tag{2.12}$$

usando (2.12) se muestra que la función de cuantiles esta dada por

$$Q(u) = F^{-1}(u) = \frac{log(\gamma) - log(-log(u - e^{-\gamma}(u-1)))}{\lambda} \tag{2.13}$$

donde u tiene una distribución uniforme $U(0,1)$.

Finalmente, la media y la varianza de T, están dadas por:

$$\begin{aligned} E(T) &= \frac{\gamma}{\lambda(e^{-\gamma}-1)} F_{2,2}([1,1],[2,2],\gamma), \\ V(T) &= \frac{\gamma}{\lambda^2(e^{-\gamma}-1)}\left[F_{3,3}([1,1,1],[2,2,2],\gamma) - \frac{\gamma}{(e^{-\gamma}-1)} F^2_{2,2}([1,1],[2,2],\gamma)\right] \end{aligned}$$

donde $F_{p,q}(a,b,\gamma)$ es la función hipergeométrica generalizada.
El siguiente resultado muestra la forma explícita de calcular los momentos de la distribución Exponencial-Poisson.

Proposición Cancho *et al.* (2011)

$$E(T^r) = \frac{\gamma\Gamma(r+1)}{\lambda^r(e^{\gamma}-1)} F_{r+1,r+1}([1,\ldots,1],[2,\ldots,2],\gamma)$$

donde $r = 0, 1, \ldots$.

2.5. Datos censurados

Entre los datos disponibles se podrían observar casos en los que el evento de interés no se llegue a presentar, a estos se les llama datos censurados. De manera más específica:
Sea $Y \sim G$, el tiempo a observar una censura con función de distribución acumulada G y sea $T \sim F$ el tiempo a la ocurrencia del evento con función de distribución acumulada F, entonces la data observada tiene la forma:

$$(\tilde{T}, \Delta) = (\min(Y,T), I(T \leq Y)). \tag{2.14}$$

donde $\tilde{T}$ es el tiempo observado e I el indicador de censura.
La función de densidad conjunta de $(\tilde{T}, \Delta)$, asumiendo que T e Y son independientes, tiene la forma:

$$\begin{aligned} f(\tilde{t}, \delta) &= \left\{f(\tilde{t})[1-G(\tilde{t})]\right\}^{\delta} \left\{[1-F(\tilde{t})]g(\tilde{t})\right\}^{(1-\delta)} \\ &\propto f(\tilde{t})^{\delta}[1-F(\tilde{t})]^{(1-\delta)} \end{aligned} \tag{2.15}$$

donde $g(\cdot)$ y $f(\cdot)$ son las funciones de densidad de Y y T, respectivamente.

2.6. Estimador de Kaplan - Meier

Supongamos que observamos una muestra aleatoria de datos sujetos a censura por la derecha $(\tilde{T}_1, \Delta_1), (\tilde{T}_2, \Delta_2), \ldots, (\tilde{T}_n, \Delta_n)$. Condicionado en la muestra, sean $t_1 < t_2 \ldots < t_k$ $(k < n)$ los tiempos de la muestra donde al menos un evento es observado y que nos permiten particionar la muestra en $k+1$ intervalos de la siguiente forma: $[t_0, t_1), [t_1, t_2) \ldots [t_k, t_{k+1})$, donde $t_0 = 0$ y $t_{k+1} = \infty$.
Para cada uno de los intervalos $[t_j, t_{j+1})$ definimos:

d_j: número de personas que experimentan evento en el tiempo t_j.

m_j de personas censuradas en el intervalo $[t_j, t_{j+1})$ en los tiempos $t_{j1}, t_{j2}, \ldots, t_{jm_j}$

n_j: número de personas en riesgo en el instante previo a t_j.

El estimador de Kaplan-Meier (Kaplan & Meier, 1958) de la función de supervivencia en el

tiempo $T = t$, está definido por:

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

2.7. Modelos de regresión paramétrica

Según autores como Kleinbaum & Klein (2015), el modelo de regresión de tiempo de falla acelerado es un modelo paramétrico en el que se asume que el logaritmo del tiempo puede ser expresado por

$$log(T) \quad = \quad \boldsymbol{\beta'X} + \sigma\epsilon, \tag{2.16}$$

donde $\boldsymbol{X} = (X_1, \ldots, X_k)^\top$ es un vector de covariables, β es el vector de coeficientes, ϵ una variable aleatoria del error y σ el parámetro de escala.
En función de (2.16) podemos definir la función de supervivencia basal

$$S_0(t) = S(t|x = 0) = P(e^{\sigma\epsilon} > t),$$

Entonces, una implicancia del modelo es lo siguiente:

$$S(t|X = x) = P(e^{\boldsymbol{\beta'x}+\sigma\epsilon} > t) = P(e^{\sigma\epsilon} > e^{-\boldsymbol{\beta'x}}) = S(t|x). \tag{2.17}$$

La ecuación (2.17) implica que el tiempo al evento acelera si $e^{-\boldsymbol{\beta'x}}$ es menor a 1 y disminuye si este es mayor a 1.
A partir de (2.16) y usando el caso particular de $\sigma = 1$ y que ϵ tiene la distribución de valores extremos estándar

$$f(\epsilon) = e^{\epsilon - e^{\epsilon}}, \epsilon > 0$$

entonces

$$f(t|x) = e^{log(t) - \boldsymbol{\beta'x} - e^{log(t)}} \frac{1}{t} = e^{-\boldsymbol{\beta'x}} e^{-te^{-\boldsymbol{\beta'x}}}, t > 0,$$

y por lo tanto

$$T|x \sim Exp(e^{-\boldsymbol{\beta'x}}). \tag{2.18}$$

A la ecuación descrita por (2.18) se le denomina el **modelo de regresión exponencial**.

2.8. El modelo de mixtura

Tal como lo propone Boag (1949), el enfoque clásico para el estudio de datos de supervivencia de larga duración es el modelo de mixtura. Además este modelo ha sido estudiado por muchos autores, entre ellos Yakovlev & Tsodikov (1993), Tsodikov (1998), y Chen *et al.* (1999). En este modelo se asume que la población de unidades de estudio está conformada por dos grupos, inmunes y susceptibles al evento de interés. Boag (1949) sugiere que el modelo es el que se encuentra descrita por la Figura 2.4.

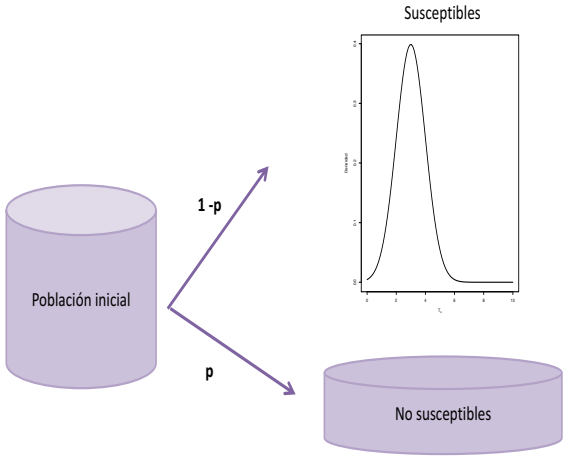

Figura 2.4: Modelo de mixtura

Según el modelo de mixtura, la v.a del tiempo a la ocurrencia del evento es de la forma:

$$T = \begin{cases} \infty \\ T_s \end{cases} \tag{2.19}$$

donde $T_s \sim F_s$ es el tiempo a la ocurrencia del evento entre susceptibles con f.d.a F_s y el tiempo entre no susceptibles se considera ∞. Entonces

$$S(t) = P(T > t) = p_0 + (1 - p_0)S_s(t), \tag{2.20}$$

donde p_0 es la proporción de unidades inmunes al evento de interés (o curados) y $S_s(\cdot)$ corresponde a la función de supervivencia de las unidades susceptibles al evento de interés. Nótese que

$$\begin{aligned} \lim_{t\to\infty} S(t) &= \lim_{t\to\infty} [p_0 + (1 - p_0)S_s(t)] \\ &= p_0 + (1 - p_0) \lim_{t\to\infty} S_s(t) \\ &= p_0 + (1 - p_0) \\ &= p_0 \end{aligned}$$

por lo tanto, $S(\cdot)$ es una función de supervivencia extendida.

Capítulo 3

El modelo de larga duración Exponencial-Poisson

En este capítulo construimos un modelo con fracción de cura y donde el tiempo a la ocurrencia del evento de interes, entre los suceptibles, es modelada mediante la distribución Exponencial-Poisson desarrollado en la sección 2.3.

3.1. Modelo

Sea M una variable aleatoria entera no negativa y latente que denota el número de factores de riesgo latentes que podrían originar la ocurrencia del evento de interés. Asumiremos que M sigue una distribución de Poisson con parámetro γ denotado por $M \sim P(\gamma)$. Adicionalmente, asumiremos que dado M, las variables aleatorias que denotan los factores de riesgo latentes, $\{Z_i\}_{i=1}^{M}$, son independientes e idénticamente distribuídas con distribución exponencial con parámetro λ.
Finalmente, asumiremos que los $\{Z_i\}_{i=1}^{M}$ son independientes de M, entonces definiremos el tiempo, T, hasta la ocurrencia del evento de interés como el tiempo que corresponde a la primera activación (Cooner *et al.* (2007)):

$$T = \begin{cases} \infty, & \text{si } M = 0, \\ \min\{ Z_1, \ldots, Z_M \}, & \text{si } M \geq 1. \end{cases} \tag{3.1}$$

Bajo esta definición, la función de supervivencia de T, en el tiempo $T = t$, esta dada por:

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(M = 0)P(T > t|M = 0) + \sum_{m=1}^{\infty} P(M = m)P(T > t|M = m) \\ &= e^{-\gamma}P(\infty > t|M = 0) + \sum_{m=1}^{\infty} e^{-\gamma}\frac{\gamma^m}{m!}P(\min\{ Z_1, \ldots, Z_M \} > t|M = m) \\ &= e^{-\gamma}1 + \sum_{m=1}^{\infty} e^{-\gamma}\frac{\gamma^m}{m!}P(\min\{ Z_1, \ldots, Z_m \} > t) \\ &= e^{-\gamma} + \sum_{m=1}^{\infty} e^{-\gamma}\frac{\gamma^m}{m!}e^{-m\lambda t} = e^{-\gamma} + e^{-\gamma}\sum_{m=1}^{\infty}\frac{(\gamma e^{-\lambda t})^m}{m!} \\ &= e^{-\gamma} + e^{-\gamma}(e^{\gamma e^{-\lambda t}} - 1) = e^{-\gamma} + (1 - e^{-\gamma})\,\frac{1 - e^{\gamma e^{-\lambda t}}}{1 - e^{\gamma}} \\ &= p_0 + (1 - p_0)S_s(t), \text{ para } t > 0 \end{aligned} \tag{3.2}$$

donde

$$p_0 = P(M = 0) = e^{-\gamma}$$

corresponde a la proporción de las unidades que no son susceptibles al evento de interés. Adicionalmente, notemos que

$$S_s(t) = \frac{1 - e^{\gamma e^{-\lambda t}}}{1 - e^{\gamma}} \ , \ t > 0 \tag{3.3}$$

corresponde a la función de supervivencia de las unidades que sí son susceptibles al evento de interés. Entonces, según lo presentado en el capítulo anterior se tiene

$$T_s \sim EP(\gamma, \lambda)$$

es decir, T_s tiene una distribución Exponencial-Poisson. Como consecuencia la función de densidad de T está dada por

$$f(t) = -\lambda e^{-\lambda t} p_0^{1-e^{-\lambda t}} log(p_0) \ , \ t > 0 \tag{3.4}$$

La Figura 3.1 muestra las funciones de densidad y supervivencia de T. De manera específica, en el lado izquierdo de la Figura 3.1 observamos la función de densidad del modelo de larga duración Exponencial-Poisson que es decreciente y tiende a cero conforme $t \to \infty$. En el lado derecho se muestra que la función de supervivencia alcanza un límite superior más alto conforme aumenta el valor de la fracción de cura, es decir, converge a uno.

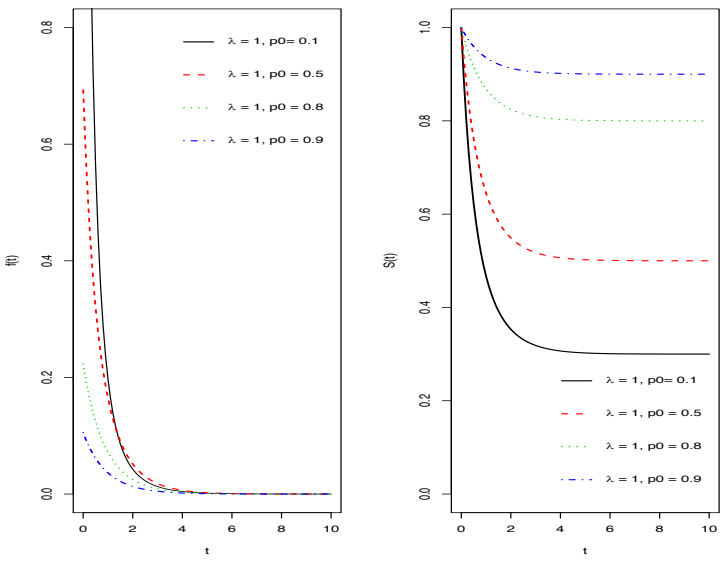

Figura 3.1: Densidad y función de supervivencia del modelo de larga duración Exponencial-Poisson

La función de riesgo esta dada por

$$h(t) = \frac{f(t)}{S(t)} = -\lambda e^{-\lambda t} log(p_0) \ , \ t > 0 \tag{3.5}$$

donde $\lambda > 0$ y $p_0 \in (0, 1)$.
La Figura 3.2 presenta la función de riesgo instantáneo del modelo la cual es decreciente y a medida que el valor de p_0 disminuye, lo hace con mayor velocidad. A medida que el tiempo aumenta, la función de riesgo instantáneo converge a cero.

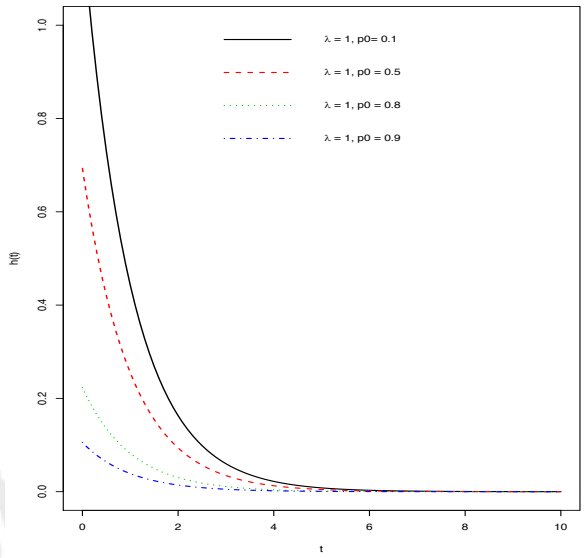

Figura 3.2: Función de riesgo instántaneo del modelo de larga duración Exponencial-Poisson

Al modelo definido por(3.1) se le denomina el modelo de larga duración Exponencial-Poisson (EP).
Una pregunta fundamental es si algunas covariables están asociadas a la proporción de unidades inmunes al evento de interés (p_0). En específico, sea $X = (X_1, X_2, ..., X_k)$ un vector de k covariables las cuales se relacionarán con la proporción de unidades inmunes al evento de interés a través de la regresión logística (Farewell, 1977):

$$p_{0i} = \frac{\exp(\boldsymbol{x}_i{}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i{}^\top \boldsymbol{\beta})}, \; i = 1, 2, \ldots, n, \tag{3.6}$$

donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ es el vector de coeficientes asociado con cada covariable.

3.2. Estructura de datos

Sea $Y \sim G$, el tiempo a observar una censura con función de distribución acumulada G y sea $T \sim F$ el tiempo a la ocurrencia del evento con función de distribución acumulada dada por uno menos la función de supervivencia definida en (1.2), entonces la data observada tiene la forma $(\tilde{T} = Y \wedge T, \delta = I(T \leq Y), \mathbf{X})$. Entonces para n observaciones $(\tilde{t}_1, \delta_1, X_1), \ldots, (\tilde{t}_n, \delta_n, X_n)$, la función de verosimilitud esta dada por:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &\propto \prod_{i=1}^{n} f(\tilde{t}_i, \boldsymbol{\theta})^{\delta_i} S(\tilde{t}_i, \boldsymbol{\theta})^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} (-\lambda e^{-\lambda t} p_{0i}^{1-e^{-\lambda t}} \log(p_{0i}))^{\delta_i} (p_{0i} + p_{0i}(p_{0i}^{-e^{-\lambda t}-1}))^{1-\delta_i} \end{aligned} \tag{3.7}$$

donde $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta}) \in \Theta = \mathbb{R}^+ \times \mathbb{R}^k$, es el vector de parámetros del modelo.

3.3. Estimación e inferencia

La función de log-verosimilitud viene dada por

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \delta_i(log(-\lambda[\boldsymbol{x}_i\beta - log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})]) + (1-e^{-\lambda t})[\boldsymbol{x}_i\boldsymbol{\beta} - log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})] - \lambda t) \\ &\quad ... + (1-\delta_i)((1-e^{-\lambda t})[\boldsymbol{x}_i\boldsymbol{\beta} - log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})]) \end{aligned} \tag{3.8}$$

entonces el estimador de máxima verosimilitud, $\hat{\theta}$, es aquel valor que maximiza la función $l(\cdot)$ en función de los parámetros λ y $\boldsymbol{\beta}$. Es decir

$$\hat{\theta} = \arg_{(\lambda,\beta)}\text{máx } \ell(.)$$

Las funciones de score para los parámetros son las siguientes:

$$\frac{\partial \ell(\theta)}{\partial \lambda} = \sum_{i=1}^{n} \left\{ \frac{(tlog(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}) - \boldsymbol{x}_i\boldsymbol{\beta}t)\lambda e^{\lambda t} - \delta_i\lambda t + \delta_i}{\lambda} \right\} \tag{3.9}$$

$$\begin{aligned}\frac{\partial \ell(\theta)}{\partial \beta_0} &= \sum_{i=1}^{n} \delta_i \left(\left(\frac{1}{(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})(\boldsymbol{x}_i\boldsymbol{\beta} - log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})} \right) + (1-e^{-\lambda t}) \left(\frac{1}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} \right) \right) \\ &\quad ... + (1-\delta_i) \left((1-e^{-\lambda t}) \left(\frac{1}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} \right) \right) \end{aligned} \tag{3.10}$$

$$\begin{aligned}\frac{\partial \ell(\theta)}{\partial \beta_j} &= \sum_{i=1}^{n} \delta_i \left(\left(\frac{x_{ij}}{(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})(\boldsymbol{x}_i\boldsymbol{\beta} - log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})} \right) + (1-e^{-\lambda t}) \left(\frac{x_{ij}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} \right) \right) \\ &\quad ... + (1-\delta_i) \left((1-e^{-\lambda t}) \left(\frac{x_{ij}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} \right) \right) \end{aligned} \tag{3.11}$$

para $j = 0, \ldots, k$. Adicionalmente, para obtener de la matriz Hessiana calculamos la segunda

derivada:

$$\frac{\partial \ell(\theta)}{\partial^2 \lambda} = \sum_{i=1}^{n} \left\{ -\frac{e^{-\lambda t}(\delta_i e^{\lambda t} + (-t^2 \log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}) + \boldsymbol{x}_i\boldsymbol{\beta} t^2)\lambda^2)}{\lambda^2} \right\} \tag{3.12}$$

$$\frac{\partial \ell(\theta)}{\partial^2 \beta_0} = \sum_{i=1}^{n} \left\{ \delta_i \left(-\frac{(e^{\boldsymbol{x}_i\boldsymbol{\beta}}(-log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}) + \boldsymbol{x}_i\boldsymbol{\beta}) + (1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})\frac{1}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}}{(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})^2(\boldsymbol{x}_i\boldsymbol{\beta} - log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}))^2} - \frac{(1-e^{-\lambda t})e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})^2} \right) \right.$$
$$\left. ... -(1-\delta_i)\frac{(1-e^{-\lambda t})e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})^2} \right\} \tag{3.13}$$

$$\frac{\partial \ell(\theta)}{\partial^2 \beta_j} = \sum_{i=1}^{n} \left\{ \delta_i \left(-\frac{x_j(x_j e^{x\beta}(-log(1+e^{x\beta}) + x\beta) + (1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})\frac{x_j}{1+e^{x\beta}}}{(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})^2(x\beta - log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}))^2} - \frac{(1-e^{-\lambda t})x_{ij}^2 e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{(1+e^{x\beta})^2} \right) \right.$$
$$\left. ... -(1-\delta_i)\frac{(1-e^{-\lambda t})x_j^2 e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})^2} \right\} \tag{3.14}$$

$$\frac{\partial \ell(\theta)}{\partial \lambda \beta_0} = \sum_{i=1}^{n} \frac{t e^{-\lambda t}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} \tag{3.15}$$

$$\frac{\partial \ell(\theta)}{\partial \lambda \beta_j} = \sum_{i=1}^{n} \frac{t x_{ij} e^{-\lambda t}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} \tag{3.16}$$

para $j = 0, \ldots, k$.

La optimización se realiza usando el método de Newton Raphson. De manera más específica, si $\theta^{(k)}$ es la estimación de θ en la iteración k, entonces la estimación en el paso $k+1$ esta dada por

$$\theta^{(k+1)} = \theta^{(k)} - H(\theta^{(k)})^{-1} \Delta \ell(\theta^{(k)}) \tag{3.17}$$

donde $\Delta\ell$ es la función de score y H es la matriz hesiana. La expresión (3.17) se usa como una ecuacion de recurrencia para dado un punto inicial generar una serie de puntos hasta converger al máximo local de $\ell(\cdot)$.

Sin embargo, en nuestro caso realizaremos la optimización haciendo uso de la función en R (R Development Core Team, 2011) **nlminb**.

Dado que estamos ante un modelo paramétrico, se cumple que asintóticamente la distribución del estimador de máxima verosimilitud es normal Hinkley & Cox (1979). En particular

$$\hat{\theta} \sim_{\text{aprox}} N\left(\theta, \frac{1}{n\hat{I}_0(\theta)}\right)$$

donde $\hat{I}_0$ es la matriz de información de Fisher observada:

$$I_0(\theta) = H(\theta)^{-1} = \begin{bmatrix} \frac{\partial \ell(\theta)}{\partial^2 \lambda} & \frac{\partial \ell(\theta)}{\partial \lambda \beta_j} \\ \frac{\partial \ell(\theta)}{\partial \lambda \beta_j} & \frac{\partial \ell(\theta)}{\beta^T \beta} \end{bmatrix}$$

Finalmente, el intervalo de confianza al $100(1-\alpha)\,\%$ para el parámetro θ_i esta dado por:

$$(\hat{\theta}_i - z_{1-\frac{\alpha}{2}}\sqrt{\hat{I}(\theta)_{ii}^{-1}}, \hat{\theta}_i + z_{1-\frac{\alpha}{2}}\sqrt{\hat{I}(\theta)_{ii}^{-1}})$$

3.4. Comparación de modelos

Uno de los criterios más utilizados para realizar las comparaciones entre modelos es el Criterio de Información de Akaike, denotado por AIC, y estudiado por Bozdogan (1987). Este criterio tiene entre sus características premiar el buen ajuste del modelo y penalizar la complejidad del mismo de acuerdo al número de parámetros utilizados. El AIC está definido de la siguiente forma

$$AIC = 2p - 2\ell(\boldsymbol{\theta})$$

donde p es el número de parámetros del modelo y $\ell((\cdot)$ es la función de log verosimilitud del modelo.

Hurvich & Tsai (1989) propusieron el estadístico AIC corregido, denotado por AICc. Este estadístico presenta una correción para tamaños de muestra bajos y está definido de la siguiente forma

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}$$

donde n es el tamaño de la muestra.

Finalmente, Schwarz *et al.* (1978) propusieron el Criterio de Información Bayesiano, denotado por BIC. Este estadístico penaliza el modelo por el número de registros de la muestra utilizada y está definido de la siguiente forma

$$BIC = \log(n)p - 2\ell(\boldsymbol{\theta}))$$

Capítulo 4

Simulación

En este capítulo se realiza un estudio de simulación para evaular la performancia del modelo, midiendo el sesgo porcentual y la cobertura de las estimaciones de los parámetros bajo distintos escenarios (usando distintos tamaños de muestra y niveles de fracción de cura).

4.1. Consideraciones para la simulación

Primero se define un valor para el parámetro λ y para los coeficientes de nuestras covariables, en este caso para β_0, β_1 y β_2. Tabla 4.1 muestra los valores de los parámetros considerados para el proceso de simulación

Cuadro 4.1: Parámetros considerados en la simulación

λ	β_0	β_1	β_2
0.8	0.5	1.2	1.1
0.8	0.5	-1.2	1.1
0.8	0.5	-1.2	-1.1

Los tamaños de muestra considerados son 250, 500 y 1000 y en cada escenario se realizaron 1000 réplicas (N=1000). Para cada escenario, se genera una muestra mediante el siguiente procedimiento:

a) Se generan las covariables X_1 y X_2, donde ambas siguen una distribución Bernoulli

$$X_{1i} \sim B(0{,}4) \ , \ X_{2i} \sim B(0{,}5) \ , \ i = 1, \ldots, n$$

b) Usando (3.6) generamos los $\{p_{0i}\}_{i=1}^{n}$ y obtenemos el parámetro $\gamma_i = \exp(-p_{0i})$ de la distribución Poisson.

c) Generamos los tiempos T_i

$$T_i = \begin{cases} \infty & \textit{con probabilidad } p_{i0} \\ T_{si} & \textit{con probabilidad } 1 - p_{i0} \end{cases}$$

donde $T_i \sim EP(\gamma_i, \lambda)$

d) Generamos los tiempos a censura

$$Y_i \sim U(0, 5) \ , \ i = 1, \ldots, n$$

e) Generamos los tiempos observados y los indicadores de censura

$$(\tilde{T}_i, \Delta_i) = (\min(Y_i, T_i), I(T_i \leq Y_i)) \ , \ i = 1, \ldots, n$$

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

Para evaluar la performancia de los estimadores, se definen los siguientes indicadores
Sesgo porcentual

$$\%Sesgo = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\hat{\theta}_i - \theta}{\theta} \right) \times 100$$

Cobertura

$$Cobertura = \frac{1}{N} \sum_{i=1}^{N} I(\theta \in 95\,\%IC(\hat{\theta}^i))$$

donde $\hat{\theta}_i$ es la estimación, via máxima verosimilitud, de θ para la muestra i-ésima

4.2. Resultados

La Tabla 4.2 muestra los resultados para tres escenarios donde se cumplen todos los supuestos de las distribuciones del modelo. En el primer escenario se considera a un 85 % de observaciones inmunes al evento de interés y en los Escenarios II y III, 75 % y 50 % respectivamente. En todos los escenarios se contempla una censura promedio del 5.2 %.

Cuadro 4.2: Sesgo porcentual y cobertura de las estimaciones: modelo bien especificado

	$n = 250$			$n = 500$			$n = 1000$		
	Sesgo (%)	Cobertura	ECM	Sesgo (%)	Cobertura	ECM	Sesgo (%)	Cobertura	ECM
Escenario I									
$\lambda = 0{,}8$	2.87	95.6	0.04	2.1	95.6	0.06	1.57	95.8	0.08
$\beta_0 = 0{,}5$	-1.75	96.4	0.08	-1.46	95.9	0.12	-1.08	95.8	0.14
$\beta_1 = 1{,}2$	5.25	95.7	0.24	3.74	96.1	0.35	2.76	96.7	0.39
$\beta_2 = 1{,}1$	2.96	95	0.16	2.38	95.9	0.24	1.84	95.5	0.28
Escenario II									
$\lambda = 0{,}8$	1.73	95.1	0.02	1.46	94.9	0.04	1.06	95	0.04
$\beta_0 = 0{,}5$	0.81	96.1	0.08	0.38	95.1	0.11	0.6	96.8	0.13
$\beta_1 = -1{,}2$	2.71	93.7	0.11	1.84	96.3	0.16	1.45	94.9	0.19
$\beta_2 = 1{,}1$	1.58	95.7	0.10	1.14	95.8	0.15	0.7	95.9	0.18
Escenario III									
$\lambda = 0{,}8$	-0.52	93.4	0.02	-0.04	94.5	0.03	0.06	94.6	0.03
$\beta_0 = 0{,}5$	0.1	95.7	0.06	-0.15	95.6	0.09	-0.31	95.6	0.11
$\beta_1 = -1{,}2$	2.86	96	0.11	1.71	93.7	0.16	1.03	94.3	0.18
$\beta_2 = -1{,}1$	2.23	95.7	0.10	1.57	95.5	0.15	1.08	95.3	0.17

En tamaños de muestra pequeños como $n = 250$, los estimadores son un poco sesgados

(sesgo porcentual alrededor de 3 %), por lo que se verá una mayor eficacia en muestras más grandes dado que también tendremos más casos donde el evento de interés se presente. Debemos considerar que para el Escenario I, si bien se observa un sesgo más elevado para $n = 250$ tal como se mencionó anteriormente pero esto sucede porque se tiene una censura alrededor de 90 % entre casos que salieron del estudio y aquellos que fueron inmunes al evento (ambos casos de censura por la derecha). Para otros tamaños de muestra superiores a $n = 250$ se observa que los estimadores si son insesgados inclusive en el Escenario I donde tenemos un alto porcentaje de censura como se explicó. En todos los casos podemos obtener la matriz hessiana y por tanto los intervalos para determinar la cobertura. Para todos los parametros ésta está alrededor de 95 %.

La Tabla 5.1 nuestra un escenario donde no se cumple el supuesto de que las variables latentes siguen una distribución exponencial. En particular, se genero $Z_i \sim Weibull(1/\lambda, 0{,}5)$ y se observo que el modelo no logra estimar correctamente el parámetro λ. Esto sugiere que al no cumplir el supuesto de que nuestra variable Z_i sigue una distribución exponencial, el modelo no tendrá una correcta performance. Notemos sin embargo, que los coeficientes de regresión son correctamente recuperados en todos los casos.

Cuadro 4.3: Sesgo porcentual y cobertura bajo la incorrecta especificación del modelo

	$n = 250$		$n = 500$		$n = 1000$	
	Sesgo porcentual	Cobertura	Sesgo porcentual	Cobertura	Sesgo porcentual	Cobertura
Escenario 4^A						
$\lambda = 0{,}8$	160.48	0	159.62	0	158.81	0
$\beta_0 = 0{,}5$	1.09	94.6	0.71	92.9	0.58	94.5
$\beta_1 = -1{,}2$	3.56	92.5	2.98	94.6	2.8	93.2
$\beta_2 = 1{,}1$	8.01	93.5	6.09	93.5	5.16	95.3
^{A}Escenario donde el tiempo fue generado por una distribución de Weibull						

Capítulo 5

Aplicación

En este capítulo se analizaran datos de un conjunto de clientes de una entidad financiera observados por 48 meses. Se consideraron los clientes cuyas solicitudes de credito habian caido en default entre Marzo y Junio del 2014. Esto quiere decir que se atrasó por más de dos meses en sus pagos. Solo se consideraron clientes que tenian una deuda mayor a S/1 000 soles. Las variables de interes son el tiempo hasta que el cliente cancele su deuda y la proporción de clientes inmunes a este evento (morosos).

Para el estudio del primer fenómeno, los tiempos censurados de la muestra serán aquellas personas cuya deuda haya sido condonada ya que esto sucede en situaciones en las que el banco considera imposible recuperar la deuda como por ejemplo si ocurre la muerte del cliente.

Adicionalmente, se evaluara mediante una regresión logistica multiple, si existe evidencia de alguna variable asociada con la probabilidad de ser inmnue (moroso). Entre las variables disponibles tenemos el radio entre el total de deuda entre y el total de desembolso, sexo del cliente, estado civil, antiguedad en el sistema financiero, edad, tipo de trabajador (independiente o no), tipo de bloqueos (castigo, judicial, refinanciado, judicial), plazo del prestamo, número de empresas financieras en las que el cliente tiene un producto vista a tres, seis y doce meses hacia atrás, deuda en otros productos, entre otras.

5.1. Descripción de los clientes

Entre las personas en el estudio, el cociente de la deuda entre el monto desembolsado es menor en aquellos que si llegan a cancelar la deuda en comparación con los que no lo hacen (63 % vs. 79 %). Cabe resaltar que los montos otorgados a los clientes siempre se dan basados en su capacidad de pago por lo que el perfil de un cliente al que se le aprobó un crédito de S/20 000 soles es mucho mejor de uno que se le aprobó uno de S/2 000 soles. Respecto a los bloqueos que pueda tener el crédito del cliente, se destaca el de refinanciado y se observa un menor porcentaje de créditos refinanciados en aquellos que cancelan su deuda en comparación con los que no lo hacen (3.10 % vs. 8.34 %). Adicionalmente, podemos observar que el porcentaje de créditos con un plazo de 12 meses es mayor en aquellos que cancelan su deuda en comparación con los que no lo hicieron (4.54 % vs. 1.63 %) y de igual manera con créditos con plazo de 36 meses (19.5 % vs. 11.26 %). A diferencia de estos dos plazos, el porcentaje de créditos con un plazo de 48 meses es menor en aquellos que cancelan su deuda en comparación con los que no lo hicieron (28.5 % vs. 39.51 %). Por otro lado, para los plazos 36 y 60 no se observa una estructura distinta entre quienes cancelan o no su deuda (Tabla

5.1).

Adicionalmente tenemos tres variables que nos indican con cuántas entidades financieras ha tenido algún producto el cliente en los últimos tres, seis y doce meses, respectivamente y tampoco se observa una estructura distinta (Tabla 5.2). Se observo que la distancia desde que se realizó el desembolso al default es mayor, en promedio, en aquellos que cancelan la deuda en comparación con los que no lo hacen (20 vs. 16).

También se tiene una variable para ver cuantos meses durante los últimos 48 meses el cliente ha tenido algún producto con algún banco y se observa que en el promedio de los meses es mayor en aquellos que cancelan a comparación con lo que no (35.11 % vs 33.28 %). Se observa que el promedio de dias de atraso es menor en aquellos que cancelan la deuda sobre los que no lo hicieron (76.92 % vs. 85.22 %).

Finalmente, se observa que la media de edad es menor en aquellos que no cancelaron la deuda en comparación con los que si lo hacieron (34 vs. 33 años). El porcentaje de clientes casados es menor en clientes que cancelaron su deuda en comparación con los que si lo hacieron (80.88 % vs. 85.59 %).

Cuadro 5.1: Características de la muestra (Parte A)

Variable	No evento 7905 (86 %)	Evento 1323 (14 %)	Todos
Ratio Deuda/Desembolso: Media (DS)	0.79 (0.23)	0.63 (0.25)	0.77 (0.24)
Bloqueo			
Sin bloqueo	6169 (78.04 %)	1088 (82.24 %)	7257 (78.64 %)
Castigada	304 (3.85 %)	97 (7.33 %)	401 (4.35 %)
Prejudicial	748 (9.46 %)	95 (7.18 %)	843 (9.14 %)
Otros	24 (0.30 %)	2 (0.15 %)	26 (0.28 %)
Refinanciado	651 (8.34 %)	49 (3.10 %)	700 (7.59 %)
Judicial	1 (0.01 %)	0 (0.00 %)	1 (0.01 %)
Refinanciamiento			
No	7246 (91.66 %)	1282 (96.90 %)	8528 (92.41 %)
Si	659 (8.34 %)	41 (3.10 %)	700 (7.59 %)
Plazo menor o igual a 12 meses			
No	7776 (98.37 %)	1263 (95.46 %)	9039 (97.95 %)
Si	129 (1.63 %)	60 (4.54 %)	189 (2.05 %)
Plazo menor o igual a 24 meses			
No	7015 (88.74 %)	1065 (80.50 %)	8080 (87.56 %)
Si	890 (11.26 %)	258 (19.50 %)	1148 (12.44 %)
Plazo menor o igual a 36 meses			
No	4200 (53.13 %)	706 (53.36 %)	4906 (53.16 %)
Si	3705 (46.87 %)	617 (46.64 %)	4322 (46.84 %)
Plazo menor o igual a 48 meses			
No	4782(60.49 %)	946(71.50 %)	5728(62.07 %)
Si	3123(39.51 %)	377 (28.50 %)	3500 (37.93 %)
Plazo menor o igual a 60 meses			
No	7847 (99.27 %)	1312 (99.17 %)	9159 (99.25 %)
Si	58 (0.73 %)	11 (0.83 %)	69 (0.75 %)

Cuadro 5.2: Características de la muestra (Parte B)

Variable	No evento 7905 (86 %)	Evento 1323 (14 %)	Todos
Max Empresas reportadas*: Media (DE)	3(1)	2(1)	3 (1)
Max Empresas reportadas**: Media (DE)	3(1)	2(1)	3 (1)
Max Empresas reportadas***: Media (DE)	3(1)	3(1)	3 (1)
Distancia de la cosecha al default: Media (DE)	16.16(8.59)	19.91 (9.83)	16.70 (8.87)
Antiguedad SF ****: Media (DE)	33.28 (13.82)	35.11 (12.98)	33.54 (13.71)
Máximo días de atraso: Media (DE) ****	85.22 (52.01)	76.92 (45.74)	84.03 (51.24)
Tenencia producto hipotecario			
No	7838 (99.15 %)	1299 (98.19 %)	9137 (99.01 %)
Si	67 (0.85 %)	24 (1.81 %)	91 (0.99 %)
Tenencia tarjeta de crédito en el último mes			
No	1850 (23.40 %)	387 (29.25 %)	2237 (24.24 %)
Si	6055 (76.60 %)	936 (70.75 %)	6991 (75.76 %)
Edad: Media (DE)	33 (9)	34 (10)	34 (9)
Casado			
No	6652 (84.15 %)	1060 (80.12 %)	7712 (83.57 %)
Si	1253 (15.85 %)	263 (19.88 %)	1516 (16.43 %)
Dependiente			
No	120 (1.52 %)	15 (1.13 %)	135 (1.46 %)
Si	7785 (98.48 %)	1308 (98.87 %)	9093 (98.54 %)
Sexo			
Femenino	1139 (14.41 %)	253 (19.12 %)	1392 (15.08 %)
Masculino	6766 (85.59 %)	1070 (80.88 %)	7836 (84.92 %)

*En los últimos 3 meses
**En los últimos 6 meses
***En los últimos 12 meses
****SF=Sistema Financiero. En los últimos 48 meses

5.2. Modelo final

Tabla 5.3 muestra las variables, su odds ratio, intervalo de confianza al 95 % y el valor de p asociado. Las covariables seleccionadas fueron el cociente de deuda entre el monto desembolsado (en porcentaje), tener tarjeta de crédito en el mes pasado, edad, estado civil (casado o no), ser independiente, sexo, tener el crédito refinanciado, tener un plazo de 36 meses, antiguedad en el sistema financiero, máximo días de atraso, tener un plazo de 48 meses, número máximo de empresas reportadas y tener un crédito hipotecario en los últimos 12 meses.

Podemos observar que al tener un cociente de deuda más elevado aumentan las chances en un 3 % de no pagar la deuda (OR: 1.03; 95 % IC: 1.03-1.04). Igualmente, al tener una tarjeta de crédito y por tanto más deuda, aumentan las chances de no pagar la deuda en un 2 % (OR: 1.02; 95 % IC: 0.78-1.32). Si se tiene más días de atraso en el historial o el cliente trabajo con más bancos, también aumentan las chances de no cancelar su deuda en 1 % (OR: 1.01; 95 % IC: 1.00-1.01) y 13 % (OR: 1.13; 95 % IC: 1.07-1.20), respectivamente. Respecto a los plazos que tienen los créditos, se tiene el indicador de plazo 36 y plazo 48 para poder ser más precisos en estas subpoblaciones ya que al tener alguno de estos dos plazos aumentan las

chances de no pagar la deuda en un 66 % (OR: 1.66; 95 % IC: 1.38-2.01) y 161 % (OR: 2.61; 95 % IC: 2.11-2.31), respectivamente. Igualmente, el ser independiente aumenta las chances de no cancelar la deuda en un 11 % (OR:1.11; 95 % IC: 0.55- 2.25) ya que un trabajador dependiente suelen tener un sueldo variable (no fijo), el mismo escenario se presenta al ser hombre (OR: 1.26; 95 % IC: 0.94-1.68).

Por otro lado, tenemos variables que al aumentar su valor, disminuyen las chances de no cancelar la deuda como lo son el tener el crédito refinanciado (OR: 0.49; 95 % IC: 0.29-0.82) ya que esto implica haber realizado un acuerdo con el banco para poder pagar su deuda de acuerdo a un calendario de pagos que se acomode a sus capacidad crediticia. Lo mismo sucede al tener más antiguedad en el sistema financiero (OR: 0.99; 95 % IC: 0.98-0.99) y al tener un producto hipotecario pues implica tener un perfil crediticio sobresaliente ya que el banco para dar un crédito de tal magnitud es muy conservador (OR: 0.27; 95 % IC: 0.14-0.52).

Observamos también que da igual la edad que tenga el cliente pues no hay una relación entre ella y no cancelar la deuda (OR: 1.00; 95 % IC: 1.00-1.02).

Respecto al valor de λ o riesgo instantáneo de ocurrencia del evento se observa que es 0.01 (95 % IC: 0.01-0.02) que es de acuerdo al esperado por el alto porcentaje de censura en la población.

Por otro lado, gracias a las covariables podemos identificar correctamente el valor de los curados o inmunes al evento el cual representa un 66 % de la población. Figura 5.1 muestra el estimador de Kaplan-Meier y este es comparado con con la estimación obtenida por el modelo Exponencial - Poisson de larga duración sin covariables.

Figura 5.2 muestre el estimador de Kaplan-Meier de la función de supervivencia y el modelo Exponencial-Poisson de larga duración para las personas casadas de la muestra.

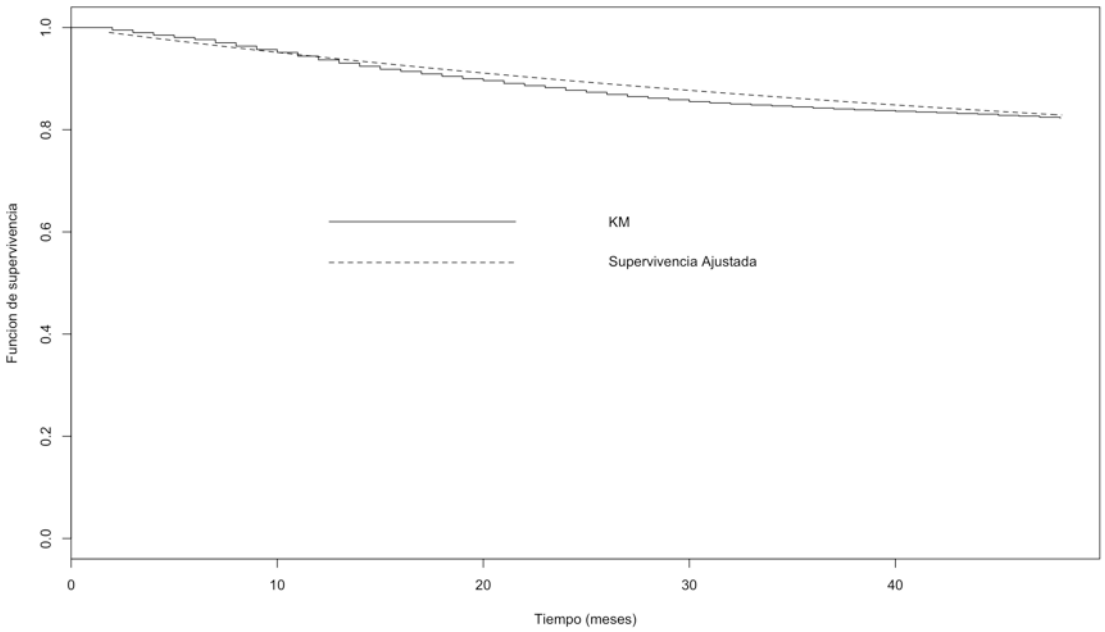

Figura 5.1: Función de supervivencia sin covariables: Kaplan-Meier vs. Exponencial-Poisson de larga duración

Cuadro 5.3: Componente de regresión logístico para el modelo de larga duración Exponencial - Poisson para evaluar factores asociados con el no pagar su deuda

Variable	β	Error estándar	p-value	OR	OR IC 95 %
Ratio de deuda (%)	0.03	0.002	<0.001	1.03	[1.03,1.04]
Tenencia TC	0.02	0.133	0.898	1.02	[0.78,1.32]
Edad	0.009	0.004	0.036	1.00	[1.00,1.02]
Casado					
No	-	-	-	-	-
Si	-0.28	0.101	0.005	0.75	[0.62, 0.92]
Independiente					
No	-	-	-	-	-
Si	0.11	0.359	0.762	1.11	[0.55,2.25]
Sexo					
Femenino	-	-	-	-	-
Masculino	0.23	0.148	0.120	1.26	[0.94,1.68]
Refinanciado					
No	-	-	-	-	-
Si	-0.71	0.2644	0.0076	0.49	[0.29,0.82]
Plazo 36 meses					
No	-	-	-	-	-
Si	0.51	0.097	<0.001	1.66	[1.38,2.01]
Antiguedad en el Sistema Financiero	-0.009	0.003	0.015	0.99	[0.98,0.99]
Máximo días de atraso	0.006	0.001	<0.001	1.01	[1.00,1.01]
Plazo 48 meses					
No	-	-	-	-	-
Si	0.96	0.1071	<0.001	2.61	[2.11,3.21]
Max. empresas reportadas	0.12	0.030	<0.001	1.13	[1.07,1.20]
Tenencia Hipotecario					
No	-	-	-	-	-
Si	-1.3	0.328	<0.001	0.27	[0.14,0.52]

5.3. Comparación de modelos

Se comparó el modelo propuesto con un modelo de larga duración Exponencial y un modelo de regresión logistica que ignora el tiempo a la ocurrencia del evento de interés. Este último modelo fue considerado pues es comunmente usado por las entidades financieras del país.

Para comparar estos modelos, se estimaron los estadísticos AIC, AICc y BIC. Se observa que el AIC del modelo Exponencial-Poisson con covariables es menor que el del mismo modelo sin covariables (17165 vs. -16271.32) y mayor que el modelo de larga duración Exponencial(Tabla 5.4). Sobre los valores de los coeficientes de regresión, se oberva que tanto entre el modelo de larga duración Exponencial-Poisson y el Exponencial, los valores se asemejan mucho.

Por otro lado, al compararlos con la regresión logística, se observa que variables como Tenencia de TC o si es refinanciado, tienen un sentido diferente. Esto se debe a que ambos modelos tienen objetivos distintos ya el modelo de larga duración Exponencial-Poisson tiene como objetivo modelar el tiempo de ocurrencia del evento y el otro la probabidad de que el

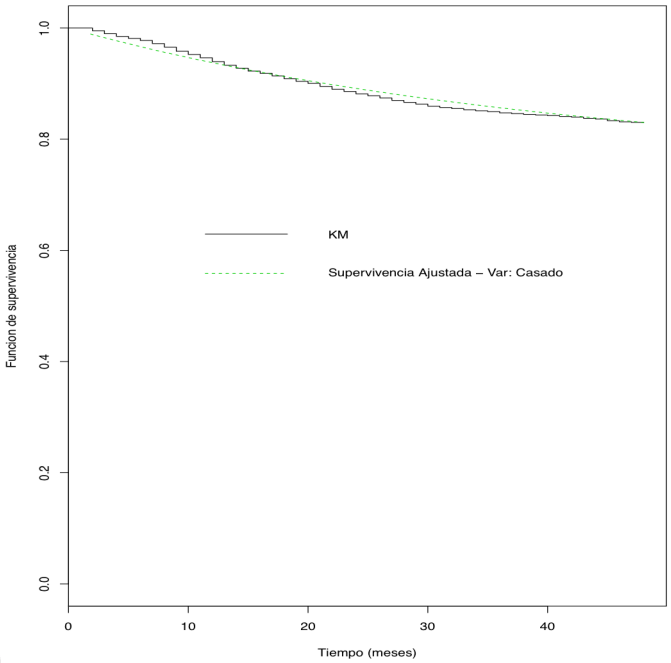

Figura 5.2: Función de supervivencia con una covariable: Ser casado

evento ocurra.

Con lo mencionado podemos comprobar que los modelos de larga duración Exponenical - Poisson se asemeja mucho al modelo de larga duración Exponencial, sin embargo, no lo supera, por lo que para la muestra utilizada este último sería el más adecuado.

Cuadro 5.4: Comparación de modelos de larga duración Exponencial-Poisson, larga duración Exponencial y Logístico

Variable	Exponencial-Poisson	Exponencial	Logístico
λ	0.01	0.02	-
Ratio de deuda (%)	0.03	0.03	0.02
Tenencia TC	0.02	0.02	-0.04
Edad	0.01	0.01	0.01
Casado			
No	-	-	-
Si	-0.28	-0.31	-0.21
Independiente			
No	-	-	-
Si	0.11	0.12	0.04
Sexo Masculino			
No	-	-	-
Si	0.23	0.25	0.32
Refinanciado			
No	-	-	-
Si	-0.71	-1.11	1.53
Plazo 36 meses			
No	-	-	-
Si	0.51	0.51	0.42
Antiguedad en el Sistema Financiero	-0.01	-0.01	-0.01
Máximo días de atraso	0.01	0.01	0.01
Plazo 48 meses			
No	-	-	-
Si	0.96	0.97	0.82
Max. empresas reportadas	0.12	0.13	0.09
Tenencia Hipotecario			
No	-	-	-
Si	-1.3	-1.39	-0.98
AIC	-16271.32	-16308.78	6868.47
AICc	-16271.27	-16308.73	6868.53
BIC	-16164.37	-16201.83	6968.24

Capítulo 6

Conclusiones

6.1. Conclusiones

En este trabajo de tesis hemos estudiado el modelo de larga duracion Exponencial-Poisson. Una de las ventajas de este modelo es que permite abordar en el análisis de supervivencia la casuística de tener una población que no es susceptible al evento de interes. Adicionalmente, se pudo estudiar que covariables podrian estar asociadas con la probabilidad de ser inmune al evento de interes a través de una regresión logística.

En el estudio de simulación se comprobó el modelo funciona bien para muestras con al menos 250 observaciones ya que se generan estimadores insesgados y con alta cobertura y va mejorando conforme el tamaño de la muestra aumenta. Observamos también que dado que el modelo es parámetrico, una violación del supuesto genera estimaciones sesgadas.

Finalmente, se aplicó el modelo sobre una población de clientes con un crédito en default (más de 60 días en atraso) donde la variable de interés es el tiempo al pago total de la deuda. Se observó que las variables ratio de deuda entre desembolso, tener tarjeta de crédito, ser independinte, ser del sexo masculino, tener un plazo de 36 meses, tener un plazo de 48 meses, máximo días de atraso y máximo de empresas reportadas aumentan el riesgo de que el cliente no cancele su deuda y ser casado, tener el crédito refinanciado y tener un crédito hipotecario, reducen las chances de cancelar la deuda.

Se observó que todas las variables dan el sentido esperado y que el modelo con covariables supera a uno sin estas y al modelo de larga duración Exponencial basándonos en el valor del AIC, es decir, a pesar de tener una mayor complejidad es superior.

6.2. Investigaciones futuras

1. Modelar la tasa de ocurrencia del evento en función de covariables. Esto para poder relacionar las características de la población con este parámetro del modelo.

2. Colocar una función más general para el tiempo. En este caso la función podría ser una Weibull-Poisson (Lu & Shi, 2012).

3. Implementación bayesiana del modelo (Chen *et al.*, 1999).

Apéndice A

Resultados teóricos

La función de log-verosimilitud del modelo sin covariables esta dada por:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n}[\,\delta_i \log(-\lambda p_{0i}^{1-e^{-\lambda t}} e^{-\lambda t} \log(p_{0i})) + (1-\delta_i)\log(p_{0i}^{1-e^{-\lambda t}})] \\
&= \sum_{i=1}^{n}[\,\delta_i(\log(-\lambda \log(p_{0i})) + \log(p_{0i}^{1-e^{-\lambda t}}) + \log(e^{-\lambda t}) + (1-\delta_i)\log(p_{0i}^{1-e^{-\lambda t}})] \\
&= \sum_{i=1}^{n}[\,\delta_i(\log(-\lambda \log(p_{0i})) + (1-e^{-\lambda t})\log(p_{0i}) - \lambda t) + (1-\delta_i)\log(p_{0i}^{1-e^{-\lambda t}})] \qquad \text{(A.1)}
\end{aligned}$$

Las funciones de score para los parámetros son:

$$\frac{\partial \ell(\theta)}{\partial p_0} = \sum_{i=1}^{n}\left[\delta_i\left(\frac{1}{\log(p_{0i})p_{0i}} + \frac{1-e^{-\lambda t}}{p_{0i}}\right) + (1-\delta_i)\left(\frac{1}{p_{0i}^{1-e^{-\lambda t}}}(1-e^{-\lambda t})p_{0i}^{-e^{-\lambda t}}\right)\right] \qquad \text{(A.2)}$$

$$\frac{\partial \ell(\theta)}{\partial \lambda} = \sum_{i=1}^{n}\left[\delta_i\left(\frac{1}{\lambda} + t\log(p_{0i})(te^{-\lambda t}) - t\right) + (1-\delta_i)(p_{0i}^{1-e^{-\lambda t}}\log(p_{0i})te^{-\lambda t})\right] \qquad \text{(A.3)}$$

Las segundas derivadas son:

$$\frac{\partial \ell(\theta)}{\partial^2 p_0} = \sum_{i=1}^{n}\left[\delta_i\left(-\frac{1}{\log(p_{0i})p_{0i}^2} - \frac{1}{\log^2(p_{0i})p_{0i}^2} - \frac{1-e^{-\lambda t}}{p_{0i}^2}\right) - (1-\delta_i)\left(\frac{1-e^{-\lambda t}}{p_{0i}^2}\right)\right] \qquad \text{(A.4)}$$

$$\frac{\partial \ell(\theta)}{\partial^2 \lambda} = \sum_{i=1}^{n}\left[\delta_i\left(-\log(p_{0i})t^3e^{-\lambda t} - \frac{1}{\lambda^2}\right) + (1-\delta_i)\log(p_{0i}) + (p_{0i}^{1-e^{-\lambda t}}\log(p_{0i})te^{-2\lambda t} - p_{0i}^{1-e^{-\lambda t}}te^{-\lambda t})\right] \qquad \text{(A.5)}$$

Apéndice B

Implementacion del codigo en R del modelo

```
#leemos la data
data<-read.csv(file.choose())

#creamos matriz con variables
X   <- as.matrix(cbind(rep(1,n),data$RATIO_DEUDA*100,
                       data$FLG_REV_1,data$EDAD,data$FLG_CASADO,
                       data$FLG_INDEPENDIENTE,
                       data$FLG_S,data$FLG_REF,
                       data$FLG_PLAZO36,
                       data$ANTIGUEDAD_RCC48,data$ATRASOMAX_1,
                       data$FLG_PLAZO48,
                       data$CTDEMPREPORTADOCLIMAX3,
                       data$FLG_HIPOTECARIO_12))

#funcion de log verosimilitud
l = function(param)
{
  lamb = exp(param[1])
  po   = as.vector(inv.logit(X%*%param[-1]))
  logF=log(-lamb*log(po))-(lamb*muestra)+(1-exp(-lamb*muestra))*log(po)
  logS=(1-exp(-lamb*muestra))*log(po)
  val= - sum(ind_cens*logF+(1-ind_cens)*logS,na.rm = T)
  return(val) ### val es el valor que dar? la funci?n Logf.
}

#Valores iniciales
#a los resultados de la regresion logistica
nind  <- 1 - ind_cens
model <- glm(nind ~ y+z+a+b+c+d+e+f+g+h+t+j+k,
             family = binomial(link=logit))

#funcion de optimizacion
```

```
library(boot)
res <- nlminb(start = c(0,model$coefficients),
              objective = l)

#obtenemos los OR
exp(res$par)

#calculo de los intervalos de confianza
library(numDeriv)

H = hessian(l,res$par)
IFO = solve(H) ### Matriz de Informacion de Fisher observada
sd = sqrt(diag(IFO))

z=qnorm(0.975)

c(res$par[3]-z*sd[3],res$par[3]+z*sd[3])
c(res$par[4]-z*sd[4],res$par[4]+z*sd[4])
c(res$par[5]-z*sd[5],res$par[5]+z*sd[5])
c(res$par[6]-z*sd[6],res$par[6]+z*sd[6])
c(res$par[7]-z*sd[7],res$par[7]+z*sd[7])
c(res$par[8]-z*sd[8],res$par[8]+z*sd[8])
c(res$par[9]-z*sd[9],res$par[9]+z*sd[9])
c(res$par[10]-z*sd[10],res$par[10]+z*sd[10])
c(res$par[11]-z*sd[11],res$par[11]+z*sd[11])
c(res$par[12]-z*sd[12],res$par[12]+z*sd[12])
c(res$par[13]-z*sd[13],res$par[13]+z*sd[13])
c(res$par[14]-z*sd[14],res$par[14]+z*sd[14])
c(res$par[15]-z*sd[15],res$par[15]+z*sd[15])
```

Apéndice C

Implementación del código en R de la simulación

```
library(numDeriv)

l = function(param)
{
  lamb = param[1]
  po = inv.logit(param[2]+param[3]*x+param[4]*y)
  logF=log(-lamb*log(po))-(lamb*muestra)+(1-exp(-lamb*muestra))*log(po)
  logS=(1-exp(-lamb*muestra))*log(po)
  val=sum(ind_cens*logF+(1-ind_cens)*logS,na.rm = T)
  return(val) ### val es el valor que dar? la funci?n Logf.
}


lambda=0.8
b0=0.5
b1=-1.2
b2=-1.1

vector_n=c(100,250,500,1000)
vector_n.tam=length(vector_n)

param_est=NULL
lim.final.lambda=NULL
lim.final.b0=NULL
lim.final.b1=NULL
lim.final.b2=NULL

nb=1000
resultados <- 1:19
names(resultados) <-
c("N","Promedio lambda", "Sesgo %lambda",
"Cobertura %lambda", "%NAN lambda",
```

```
"Promedio b0", "Sesgo %b0", "Cobertura %b0", "%NAN b0",
"Promedio b1", "Sesgo %b1", "Cobertura %b1", "%NAN b1",
"Promedio b2", "Sesgo %b2", "Cobertura %b2", "%NAN b2",
"% censura","% NAN Hes")


for (k in 1:vector_n.tam)
{
  n=vector_n[k]

  cobertura.lambda=0
  cobertura.b0=0
  cobertura.b1=0
  cobertura.b2=0

  contadornan.lambda=0
  contadornan.b0=0
  contadornan.b1=0
  contadornan.b2=0
  vector_censura_porc=NULL
  contadornan.h=0

  for (j in 1:nb)
  {
    ######covariables######
    x<-rbinom(n,1,0.4)
    y<-rbinom(n,1,0.5)
    po<-inv.logit(b0+b1*x+b2*y)
    #summary(po)
    gamma=-log(po)
    #summary(gamma)
    ##########################

    M<-rpois(n,gamma)

    c=runif(n,0,5)

    T1<-NULL
    T<-NULL
    ind_censura=NULL

    contador.censura=0
```

```
for(i in 1:n)
{
  if(M[i]>=1){
    T1[i]=min(rexp(M[i],lambda))
    T[i]=T1[i]
    if (c[i]<T1[i]) {
      ind_censura[i]=0
      contador.censura=contador.censura+1
      T[i]=c[i]
    }
    else{
      ind_censura[i]=1
      summary(ind_censura)
    }
  } else {
    T1[i]=Inf
    T[i]=c[i]
    #T[i]=Inf
    ind_censura[i]=0
  }
}

vector_censura_porc[j]=contador.censura

data=cbind(T1,c,T,ind_censura)
data=data.frame(data)

muestra = data[,3]##data[,4]
ind_cens = data[,4]##data[,5]

valor_inicial = c(0.5,0.2,0.8,0.8)

library(BB)
est_mv = BBoptim(fn = l, par = valor_inicial,
lower=c(0,-Inf,-Inf,-Inf),
upper=c(Inf,Inf,Inf,Inf),
control=list(maximize = TRUE), method = c(2, 3, 1))

temp_est_mv=NULL
temp_est_mv[1]=as.numeric(est_mv$par[1])
temp_est_mv[2]=as.numeric(est_mv$par[2])
```

```
temp_est_mv[3]=as.numeric(est_mv$par[3])
temp_est_mv[4]=as.numeric(est_mv$par[4])
temp_est_mv[5]=n


param_est=rbind(param_est,temp_est_mv)

#intervalos de confianza
H = hessian(l,est_mv$par, "Richardson")
sd=NULL

cal.hes=try(solve(-H),silent=TRUE)

if ( class(cal.hes)=="matrix") {

IFO = solve(-H) #Matriz de Informacion de Fisher observada
sd = sqrt(diag(IFO))

z=qnorm(0.975)

lim.lambda=NULL
lim.b0=NULL
lim.b1=NULL
lim.b2=NULL

lim.lambda[1]=est_mv$par[1]-z*sd[1]
lim.lambda[2]=est_mv$par[1]+z*sd[1]
lim.lambda[3]=n

lim.final.lambda=rbind(lim.final.lambda,lim.lambda)

lim.b0[1]=est_mv$par[2]-z*sd[2]
lim.b0[2]=est_mv$par[2]+z*sd[2]
lim.b0[3]=n

lim.final.b0=rbind(lim.final.b0,lim.b0)

lim.b1[1]=est_mv$par[3]-z*sd[3]
lim.b1[2]=est_mv$par[3]+z*sd[3]
lim.b1[3]=n

lim.final.b1=rbind(lim.final.b1,lim.b1)
```

```
    lim.b2[1]=est_mv$par[4]-z*sd[4]
    lim.b2[2]=est_mv$par[4]+z*sd[4]
    lim.b2[3]=n

    lim.final.b2=rbind(lim.final.b2,lim.b2)

    if (is.nan(lim.lambda[1]) || is.nan(lim.lambda[2])) {
      contadornan.lambda=contadornan.lambda+1
    } else if (lambda>=lim.lambda[1] && lambda<=lim.lambda[2]){
      cobertura.lambda=cobertura.lambda+1}

    if (is.nan(lim.b0[1]) || is.nan(lim.b0[2])) {
      contadornan.b0=contadornan.b0+1
      }else if (b0>=lim.b0[1] && b0<=lim.b0[2]){
      cobertura.b0=cobertura.b0+1}

    if (is.nan(lim.b1[1]) || is.nan(lim.b1[2])) {
      contadornan.b1=contadornan.b1+1
    }else if (b1>=lim.b1[1] && b1<=lim.b1[2]){
      cobertura.b1=cobertura.b1+1}

    if (is.nan(lim.b2[1]) || is.nan(lim.b2[2])) {
      contadornan.b2=contadornan.b2+1
    }else if (b2>=lim.b2[1] && b2<=lim.b2[2]){
      cobertura.b2=cobertura.b2+1}

    }
    else
    {
      contadornan.h=contadornan.h+1
    }

  }

  censura.porc=mean(vector_censura_porc)/n*100
  sesgo.lambda=(mean(param_est[,1])-lambda)/lambda
  sesgo.b0=(mean(param_est[,2])-b0)/b0
  sesgo.b1=(mean(param_est[,3])-b1)/b1
  sesgo.b2=(mean(param_est[,4])-b2)/b2

  ################################
```

```
  #promedios
  lambda.prom=mean(param_est[,1])
  b0.prom=mean(param_est[,2])
  b1.prom=mean(param_est[,3])
  b2.prom=mean(param_est[,4])
  #sesgo porcentual
  sesgo.lambda.porc=sesgo.lambda*100
  sesgo.b0.porc=sesgo.b0*100
  sesgo.b1.porc=sesgo.b1*100
  sesgo.b2.porc=sesgo.b2*100
  #cobertura
  cobertura.lambda.porc=cobertura.lambda/nb*100
  cobertura.b0.porc=cobertura.b0/nb*100
  cobertura.b1.porc=cobertura.b1/nb*100
  cobertura.b2.porc=cobertura.b2/nb*100
  #contador intervalos nan
  contadornan.lambda.porc=contadornan.lambda/nb*100
  contadornan.b0.porc=contadornan.b0/nb*100
  contadornan.b1.porc=contadornan.b1/nb*100
  contadornan.b2.porc=contadornan.b2/nb*100
  contadornan.h.poc=contadornan.h/nb*100

  resultados.temp=c(n,lambda.prom,
                                  sesgo.lambda.porc,
                                  cobertura.lambda.porc,
                                  contadornan.lambda.porc,
                    b0.prom,sesgo.b0.porc,
                    cobertura.b0.porc,contadornan.b0.porc,
                    b1.prom,sesgo.b1.porc,
                    cobertura.b1.porc,contadornan.b1.porc,
                    b2.prom,sesgo.b2.porc,
                    cobertura.b2.porc,contadornan.b2.porc,
                    censura.porc,contadornan.h.poc)

  resultados=rbind(resultados,resultados.temp)
}
```

Bibliografía

Ash, R. B. (1972). *Real analysis and probability*. New York, Academic Press.

Barreto-Souza, W. & Cribari-Neto, F. (2009). A generalization of the exponential-poisson distribution. *Statistics & Probability Letters*, **79**(24), 2493–2500.

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.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, **52**(3), 345–370.

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

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

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

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

Hinkley, D. V. & Cox, D. (1979). *Theoretical statistics*. Chapman and Hall/CRC.

Hosmer, D. W., Lemeshow, S. & May, S. (2011). *Applied survival analysis*. Wiley Blackwell.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**(2), 297–307.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), pp. 457–481.

Karlis, D. (2009). A note on the exponential poisson distribution: A nested em algorithm. *Computational Statistics & Data Analysis*, **53**(4), 894–899.

Kilbas, A., Saxena, R., Saigo, M. & Trujillo, J. (2012). Series representations and asymptotic expansions of extended generalized hypergeometric function. *Analytic Methods of Analysis and Differential Equations: AMADE-2009*, pages 31–59.

Kleinbaum, D. & Klein, M. (2015). *Survival Analysis*. Springer.

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

Louzada, F., Cancho, V. G. & Barriga, G. D. (2012). The poisson-exponential regression model under different latent activation schemes. *Computational & Applied Mathematics*, **31**(3), 617–632.

Lu, W. & Shi, D. (2012). A new compounding life distribution: the weibull–poisson distribution. *Journal of applied statistics*, **39**(1), 21–38.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rodrigues, G. C., Louzada, F. & Ramos, P. L. (2018). Poisson–exponential distribution: different methods of estimation. *Journal of Applied Statistics*, **45**(1), 128–144.

Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, **54**(1), 1508–1516.

Yakovlev, A. Y. & Tsodikov, A. D. (1993). A stochastic-model of hormesis. *Mathematical Biosciences*, **116**(2), 197–219.