UNIVERSIDAD NACIONAL AGRARIA

LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"MODELOS SEMIPARAMÉTRICOS DE EVENTOS RECURRENTES: CASO APLICACIÓN A PACIENTES CON CANCER DE MAMA"

Presentada por:
DIANA DEL ROCÍO REBAZA FERNÁNDEZ

TESIS PARA OPTAR EL GRADO DE MAGISTER SCIENTIAE EN

ESTADÍSTICA APLICADA

Lima – Perú
2017

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"MODELOS SEMIPARAMÉTRICOS DE EVENTOS RECURRENTES: CASO APLICACIÓN A PACIENTES CON CANCER DE MAMA"

Presentada por:

DIANA DEL ROCÍO REBAZA FERNÁNDEZ

TESIS PARA OPTAR EL GRADO DE MAGISTER SCIENTIAE EN ESTADÍSTICA APLICADA

SUSTENTADA Y APROBADA POR EL SIGUIENTE JURADO

Mg.Sc. César Higinio Menacho Chiok
PRESIDENTE

Mg.Sc. Víctor Manuel Maehara Oyata
PATROCINADOR

Mg.Sc. Grimaldo José Febres Huamán
MIEMBRO

Mg.Sc. Rino Nicanor Sotomayor Ruíz
MIEMBRO

DEDICATORIA

La presente investigación está dedicada a mis padres Carlos Rebaza Tamayo y Rosa Fernández Becerra por el amor, el cariño y la confianza que depositaron en mí, por el apoyo constante durante mi vida académica y profesional.

AGRADECIMIENTOS

Agradezco en primer lugar a Dios todo poderoso que ha estado conmigo durante el transcurso mi vida, quien me cuida y guía en la toma de decisiones.

A mis queridos padres por el apoyo constante incondicional, a mis hermanos Carlos, Miguel y Lorena por el cariño y apoyo a la distancia. De manera muy especial a mi hermana Rosita por su confianza, apoyo y cariño aún a la distancia.

A la Universidad Nacional Agraria La Molina que me acogió como estudiante de maestría y ahora formar parte de su plana docente. A sus trabajadores en especial a la secretaria del Departamento de Estadística e Informática Señora Rosa Sacsa.

Al MS. Victor Maehara Oyata por su asesoría, consejos y apoyo en la elaboración de la presente investigación.

Al Dr. Vladimir Villoslada Terrones por el apoyo con los datos de casos de pacientes con cáncer de mama.

A mi mejor amigo, colega y hermano Joao, por estar siempre en las buenas y en las malas, por sus reiteradas demostraciones de amistad para encaminar la presente investigación.

A mis amigos Rolando, Aldo, Jesús, Anita y Yency por su apoyo y amistad incondicional.

A todos y cada una de personas que contribuyeron de alguna manera en la realización de la presente investigación.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

RESUMEN

La recurrencia de un evento en un paciente es la frecuencia observada de este en un periodo de tiempo durante el seguimiento al individuo, por ejemplo hospitalizaciones sucesivas de neumonía, episodios de epilepsia, recaídas de cáncer, entre otros. Los modelos de eventos recurrentes son muy útiles para la aplicación en estos fenómenos, y la presente investigación pretende ilustrar y comparar modelos particulares de datos de eventos recurrentes sin efecto aleatorio: Andersen y Gill (A-D); Wei, Lin y Weissfeld (WLW); y, Prentice, Williams y Peterson (PWP), los cuales son modelos basados en la extensión de Cox de riesgos proporcionales, en estos modelos se asumen independencia de eventos. Otro modelo estudiado es el modelo de Fragilidad Compartida Gamma para eventos recurrentes que considera un término de fragilidad y asume que este término influye en la recurrencia de los eventos de un mismo sujeto. Para la estimación de los parámetros en los modelos sin efecto aleatorio se utilizó el método de máxima verosimilitud parcial mientras que para el modelo de fragilidad fue el método de máxima verosimilitud penalizado, el cual penaliza la función de riesgo base. Los datos usados para la aplicación de estas metodologías fue proporcionada por el médico Ginecólogo Oncólogo Dr. Vladimir Villoslada Terrones del Instituto Nacional de Enfermedades Neoplásicas (INEN). Estos datos describen un conjunto de variables relacionados al cáncer de mama en una cohorte prospectiva de 68 pacientes con diagnóstico positivo, sometidos a una cirugía mastectomía. Al procesar y analizar los resultados obtenidos, se encontró que el modelo Andersen y Gill (A-D) y Prentice, Williams y Peterson (PWP) son los que ajustan mejor a este conjunto de datos. Entre los resultados encontrados se obtuvo que los factores asociados al riesgo de recurrencia de cáncer de mama son la edad de inicio al estudio, la edad de primera menstruación (menarquia) y tipo carcinoma lobulillar. Estos modelos presentan similares resultados debido a la significancia estadística en las variables y el cumplimiento del supuesto de riesgos proporcionales.

Palabras Claves: Recurrencia cáncer de mama, Modelo A-D, Modelo PWP, Modelo WLW, verosimilitud penalizada, Modelo de Fragilidad

SUMMARY

The recurrence of an event in a patient is the observed frequency of this event over a period of time during follow-up, e.g. successive hospitalizations of pneumonia, episodes of epilepsy, relapses of cancer, among others. Recurrent event models are very useful for application in these phenomena, and the present research is intended to illustrate and compare particular models for recurrent event data without random effect: Andersen and Gill (A-G); Wei, Lin and Weissfeld (WLW); and Prentice, Williams and Peterson (PWP), which are models based on the Cox extension of proportional hazards, in these models assume independence of events. Another studied model is the Gamma Shared Fragility model that considers a term of fragility and assumes that this term influences the recurrence of the events of the same subject. For the estimation of the parameters in the models without random effect, the maximum likelihood method was used, while for the fragility model was the penalized maximum likelihood method, which penalizes the function of base risk. The data used for the application of these methodologies was provided by the physician Gynecologist Oncologist Dr. Vladimir Villoslada Terrones of the National Institute of Neoplastic Diseases (INEN, in its Spanish acronym). These data describe a set of variables related to breast cancer in a prospective cohort of 68 patients with positive diagnosis undergoing mastectomy surgery. When processing and analyzing the obtained results, we found that the model Andersen and Gill (A-G) and Prentice, Williams and Peterson (PWP) are the best fit to this data set. Besides, we found that the factors associated with risk of recurrence of breast cancer are the age of onset of the study, the age of first menstruation (menarche) and lobular carcinoma type. These models present similar results due to the statistical significance in the variables and compliance with the proportional risk assumption.

Key words: Breast cancer recurrence, Model A-G, PWP model, WLW model, penalized likelihood, Fragility model

I. INTRODUCCIÓN

Las investigaciones relacionadas a supervivencia con un único evento por unidad de estudio han sido ampliamente usadas, como por ejemplo el seguimiento de cierta enfermedad con un tratamiento hasta que ocurra el evento de interés, el cual puede ser: la muerte o mejora del paciente. Sin embargo, hoy en día los investigadores también se han preocupado por estudiar la ocurrencia de un evento en más de una vez; es decir, su recurrencia en el tiempo, por ejemplo, el número de recaídas de un paciente luego de someterse a una intervención quirúrgica, los ataques recurrentes del corazón en pacientes con enfermedad de las arterias coronarias, recurrencia de hospitalizaciones, la recurrencia de mejoras luego de someterse a tratamientos, entre otros. Es necesario recalcar que esta recurrencia no sólo suele darse en el ámbito de la medicina, también suele observarse en otros ámbitos: como las fallas de cierta máquina, la producción de artículos defectuosos, el inicio de una depresión, las denuncias por maltrato familiar, la recurrencia en robos, asaltos, deserción en la industria de telecomunicaciones (Cárdenas, 2013), entre otros casos en que se observa la frecuencia del evento en una misma unidad de estudio. Estos tipos de datos tienen como objetivo evaluar la relación de predictores con la tasa de ocurrencia, permitiendo múltiples eventos por sujeto (Kleinbaum y Klein, 2005).

Diferentes metodologías en el análisis de supervivencia han venido siendo estudiadas por varios autores, con la finalidad de obtener buenos estimadores de la función de supervivencia para el comportamiento de fenómenos recurrentes. En el ámbito biomédico, los casos más frecuentes para modelar los eventos recurrentes es cuando existe o no correlación en los tiempos de eventos recurrentes por individuo, y de acuerdo a cada caso, existen modelos específicos que se puede utilizar, y, si no se tiene en cuenta este aspecto podría obtenerse estimadores sesgados e ineficientes (Gonzales y Peña, 2004).

Se puede encontrar diversas investigaciones relacionadas a eventos recurrentes en distintas partes del mundo, las cuales se diferencian por el tipo de modelo aplicado y su realidad social. Respecto a los modelos aplicados en la presente investigación podemos mencionar algunos antecedentes a nivel internacional: Ullah et al. (2014), compararon cinco modelos de eventos recurrentes en un caso de aplicación con datos de lesiones recurrentes de la Liga Nacional de Rugby de Australia durante la temporada del 2008. Rondeau, V. (2010), presenta en su artículo varios modelos, entre ellos el modelo de fragilidad para eventos recurrentes, dentro del cual realiza la aplicación a casos de pacientes con cáncer de mama del Instituto de Bergonié al Suroeste de Francia. El autor hace uso de la estimación de máxima verosimilitud penalizadas, y la estimación de parámetros es obtenido con el algoritmo robusto Marquardt, el cual se menciona en la presente investigación. En américa latina se encontró una investigación realizada en Colombia por Cárdena, M. (2013), cuyo objetivo fue modelar el riesgo de pérdida de clientes del segmento empresarial, en la industria de telecomunicaciones, haciendo uso de un modelo de sobrevida para eventos recurrentes en presencia de un evento terminal. Así como los trabajos mencionados anteriormente, diversas investigaciones se pueden encontrar a nivel internacional. Sin embargo en el Perú no se han encontrado investigaciones relacionadas al análisis de fenómenos o eventos recurrentes, y, mediante la exposición de estas metodologías en la presente investigación, así como su aplicación a un caso con datos reales abrirá paso al interés de muchos investigadores en el área de la salud para encontrar respuestas a muchos problemas e interrogantes que encuentren en sus investigaciones, de manera que permita diseñar nuevas políticas preventivas de salud.

1.1. Objetivos

La presente investigación tiene los siguientes objetivos.

1.1.1. Objetivo principal

Comparar modelos semiparamétricos de regresión que mejor se ajusta al conjunto de datos de tiempos de recurrencia de cáncer de mama en pacientes con diagnóstico positivo del Instituto Nacional de Enfermedades Neoplásicas. Luego, como

consecuencia de este ajuste identificar los principales factores de riesgo asociado con la recurrencia.

1.1.2. Objetivos secundarios

- Determinar el número promedio de eventos recurrentes por unidad de estudio.
- Obtener y analizar las medidas descriptivas de las variables en estudio.
- Encontrar la función de riesgo y los coeficientes de regresión estimados para los modelos expuestos, Andersen y Gill (A-G); Wei, Lin y Weissfeld (WLW); y Prentice, Williams y Peterson (PWP) y Modelo de Fragilidad Compartida.
- Encontrar intervalos de confianza para los riesgos estimados.
- Analizar los residual, cabe mencionar que la presente investigación realiza la adaptación de las residuales martingalas para eventos recurrentes con el fin de evaluar su forma funcional, valores influyentes y la asunción de riesgos proporcionales el cual no se encuentran directamente desarrolladas en la bibliografía.
- Analizar los resultados encontrados.

II. REVISIÓN DE LITERATURA

2.1. Conceptos Básicos

2.1.1. Proceso de Conteo

Barbosa y Linás (2013), define el proceso de conteo como un ejemplo de proceso estocástico $\{N_t, t \geq 0\}$, si N_t representa el número total de eventos que ocurren hasta el tiempo *t*. Por lo tanto, todo proceso de conteo debe de satisfacer las siguientes condiciones:

(a) $N_t \geq 0$, para todo $t \geq 0$.

(b) Para cada $t \geq 0$, la variable N_t tiene un valor entero.

(c) Para cada $s < t$, el incremento $N_t - N_s$, es igual al número de eventos que ocurren en el intervalo de tiempo $(s,t]$.

Therneau y Grambsch (2000), define el proceso de conteo $N = \{N(t), t \geq 0\}$ como un proceso estocástico de inicio 0, cuyo comportamiento de las observaciones son continuas a la derecha y forman funciones escalonadas de altura 1.

La formulación del proceso de conteo reemplaza el par de variables $(N_i(t), Y_i(t))$, donde:

$N_i(t)$= el número de eventos observados de $[0,t]$ por unidad *i*

$$Y_i(t) = \begin{cases} 1 = \text{ unidad } i \text{ bajo observacion y en riesgo a l tiempo } t \\ 0 = \text{otro caso} \end{cases}$$

Esta formulación incluye los datos censurados por la derecha como un caso especial de proceso de conteo. Se generaliza inmediatamente a múltiples eventos y múltiples intervalos de riesgo, ampliando el alcance de los procesos más elaborados, como las no estacionarias de Poisson, Markov y multiestados, renovación modulada y procesos semi-Markov.

Se puede generalizar el proceso de conteo a datos de eventos recurrentes. Se tiene para *j*= 1,2,… y *i*= 1,2,…,*n*, definida $C_{ij} = \tau_i - S_{ij-1}$,

$$N_{ij}(t) = I\{T_{ij} \le t, T_{ij} \le C_{ij}\},$$

Y

$$Y_{ij}(t) = I\{T_{ij} \ge t, C_{ij} \ge t\},$$

Donde los procesos $N_{ij}(t)$ y $Y_{ij}(t)$ están definidas por $t \geq 0$, y se asume $N_{ij}(0) = 0$ y $Y_{ij}(0) = 1$.

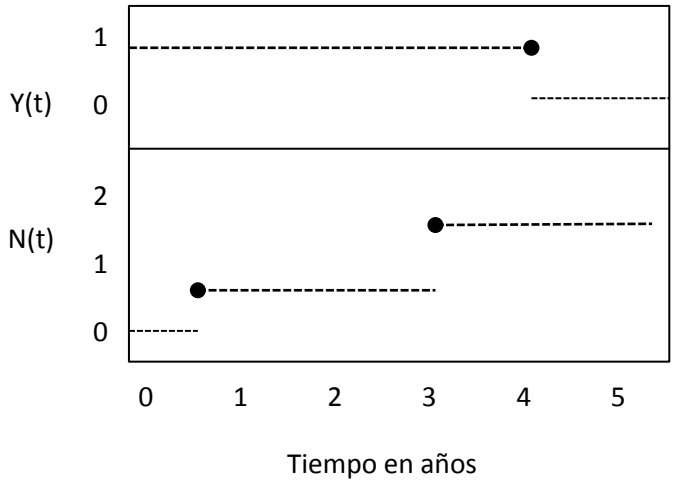


Figura 1: Ilustración del proceso N y Y. Fuente: Therneau y Grambsch (2000)

La figura 1, representa a un sujeto con dos eventos, el primer evento sucedió al año 0.5, N(0.5)=1, el segundo evento en el año 3, N(3)=2, asimismo, se encuentra en riesgo y es seguido hasta el año 4, Y(4)=1, luego de la cual deja de estar en riesgo Y(t)=0.

2.1.2. Martingalas

Karlin y Taylor (1998) definen a martingala como un proceso estocástico X_t de valor real con el conjunto de parámetros discretos o continuos. Se dice que X_t es una martingala si, $E[|X_t|] < \infty$ para todo *t*, y, si para $t_1 < t_2 < ... < t_{n+1}$, $E(X_{t_{n+1}} \mid X_{t_1} = a_1, ..., X_{t_n} = a_n) = a_n$. Las martingalas pueden ser consideradas como modelos apropiados para juegos, en el sentido de que X_t significa la cantidad de dinero que un jugador tiene en el tiempo t. Los estados de la propiedad martingala consisten en que la cantidad promedio de un jugador tendrá al tiempo t_{n+1}, dado que tiene una cantidad en el tiempo t_n, son iguales,

independientemente de su fortuna en el pasado. El proceso $X_n = Z_1 + ... + Z_n$, *n=1,2,…*, es una martingala de tiempo discreto siempre y cuando Z_i sean independientes con media cero. Del mismo modo, si X_t, $0 \leq t < \infty$ tiene incrementos independientes cuyas medias son cero, entonces X_t es una martingala de tiempo continuo.

2.1.3. Los residuales martingala

Es la diferencia entre el número observado de eventos para un individuo y el número condicionalmente esperado de eventos dado el modelo de supervivencia ajustado, tiempo de seguimiento, y el curso de las covariables observadas que varían en el tiempo. (Therneau y Grambsch, 2000)

2.1.4. Definición de Censura y Tipos

Una observación censurada es una observación incompleta; es decir contiene sólo información parcial sobre el tiempo del evento. Eso significa que el paciente es seguido durante algún tiempo, pero el evento no ocurre durante este período. Sólo se sabe que el tiempo de evento verdadero excede el tiempo de censura observado (Wienke, 2011).
La censura se clasifica de la siguiente manera:

- **Censura a la derecha**

Ocurre cuando del individuo solo se conoce que el tiempo del evento se ubica a la derecha del tiempo censurado. Klein y Moeschberger, 2003 menciona que la censura a la derecha se puede dividir en tres categorías:

- Censura tipo I: el evento se observa solamente si ocurre antes de algún tiempo preespecificado. Estos tiempos de censura pueden variar de individuo a individuo.
- Censura tipo II: sucede cuando el estudio continúa hasta la ocurrencia del evento de los primeros *r* individuos, donde r es un entero predeterminado (r <n).
- Censura aleatoria: Conocer la causa de la censura es de principal atención para evitar estimadores de supervivencia sesgados. En casos biomédicos, una de las causas de la censura aleatoria es el abandono del paciente. Si el abandono del paciente ocurre al azar y no está relacionado con el proceso de la enfermedad, tal censura no puede causar ningún sesgo en el análisis. Sin embargo, si el paciente está cerca de la muerte son más

propensos a abandonar que otros pacientes, sesgos graves pueden surgir. Otra causa de censura aleatoria es los eventos competitivos (Moore, 2016).

- **Censura a la izquierda**

Censura a la izquierda es equivalente al tiempo de entrada retardado, cuando el evento de interés ya ocurrió antes que el individuo sea observado en el estudio en el tiempo C_i (Xian, 2012).

2.1.5. Fragilidad

Duchateau y Janssen (2008) mencionan que el término fragilidad se origina en el área de medicina en la especialidad de gerontología el que se utiliza para indicar que las personas frágiles tienen un mayor riesgo de morbilidad y mortalidad.
Asimismo, la introducción de un efecto aleatorio al modelo de datos de supervivencia se debe a Beard en 1959, y el propósito de introducir el efecto era mejorar la modelización de la mortalidad en una población, el cual utilizó el factor de la longevidad en lugar de fragilidad.

Otros autores (Vaupel *et al*. 1979,1986; Hougaard,1984; Vaupel y Yashin, 1985; Hougaard, 1987), interpretan a la fragilidad como el efecto de las covariables no observadas, lo que lleva a que algunos pacientes tiendan a experimentar más eventos que otros. Asimismo, también mencionan que los modelos de fragilidad asumen que la distribución de esos efectos individuales podría ser conocida.

2.2. Modelos Clásicos de supervivencia

2.2.1. Función de Supervivencia

Considerada como una de las principales funciones probabilísticas usadas para describir estudios de supervivencia y es definida como la probabilidad de una observación de no fallar hasta un cierto tiempo *t*, es decir, la probabilidad de sobrevivir en el período de observación *t*. (Colosimo y Ruiz, 2006)

Sea $F(t)$, definida como la probabilidad que un individuo sobreviva a lo más un tiempo t:

$$F(t) = P(T \leq t) = \int_0^t f(t)dt \tag{2.1}$$

Entonces la función de supervivencia $S(t)$ está dada por:

$S(t)$=P(un individuo sobreviva un tiempo mayor a t)

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

$S(t)$ es una función no creciente del tiempo t con las propiedades:

$$S(t) = \begin{cases} 1, & t = 0 \\ 0, & t \Rightarrow \infty \end{cases} \tag{2.3}$$

Es decir, la probabilidad de sobrevivir al tiempo cero es 1 y el de sobrevivir a un tiempo infinito es cero. La función $S(t)$ también se conoce como la tasa de supervivencia acumulada que se utiliza para describir el curso de supervivencia (Berkson, 1942).

2.2.2. Función de densidad de probabilidad

Al igual que cualquier otra variable continua, el tiempo T (supervivencia) tiene una función de densidad de probabilidad definida como el límite de la probabilidad de que un individuo falle por unidad de tiempo con intervalo corto $[t,t+\Delta t)$, de ancho Δt, o simplemente la probabilidad de falla en un intervalo corto por unidad de tiempo. Esto puede ser expresado como:

$$f(t) = \frac{\lim_{\Delta t \to 0} P(t \leq T < t + \Delta t)}{\Delta t} \tag{2.4}$$

La función de densidad tiene las siguientes propiedades:

a. $f(t)$ es una función no negativa:

$$\begin{cases} f(t) \geq 0, & para\ todo\ t \geq 0 \\ = 0, & para\ t < 0 \end{cases} \tag{2.5}$$

b. En la práctica, si no hay observaciones censuradas, la probabilidad de la función de densidad $f(t)$ es estimada como la proporción de individuos que fallecen en un intervalo por unidad de ancho.

Similar a la estimación de $S(t)$, cuando las observaciones censuradas están presentes, la formula anterior no es aplicable.

La función de densidad también se conoce como la tasa de falla incondicional.

2.2.3. Función de riesgo

La función de riesgo $h(t)$ es la tasa de riesgo también denominado tasa de muerte (o falla) instantánea. La tasa de riesgo es obtenida de la probabilidad condicional que un evento ocurra en el intervalo $[t, t+\Delta t[$ dado que el evento no ocurrió aún antes del tiempo t. La tasa es obtenida dividiendo esta probabilidad condicional por el intervalo Δt (resultando en una probabilidad condicional por unidad de tiempo). La tasa de riesgo es el límite de esta tasa para Δt tendiendo cero (Duchateau y Janssen, 2008).

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid \mathrm{T} \ge \mathrm{t})}{\Delta t} = \frac{f(t)}{S(t)} \tag{2.6}$$

La función de riesgo puede ser definida en términos de función de distribución acumulada $F(t)$ y la función de densidad de probabilidad f(t):

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \tag{2.7}$$

2.2.4. Estimador no paramétrico de Kaplan–Meier:

El estimador de Kaplan-Meier de producto límite de la función de supervivencia (Kaplan y Meier, 1958), es un estimador que incorpora información de todas las observaciones disponibles, tanto de datos censurados como no censurados. Se considera la supervivencia en cualquier punto en el tiempo, como una serie definidos por la supervivencia observada y los tiempos censurados (Hosmer y Lemeshow, 1999).

Este estimador se encuentra mediante el siguiente procedimiento: sea k-tiempos de vidas tal que $t_1 < t_2 < \ldots < t_k$ con d_j muertos al tiempo t_j y n_j individuos en riesgo al tiempo t_j. Además, suponga que en el intervalo $[t_{j-1}, t)$ ocurrieron c_j observaciones censuradas entonces el estimador Kaplan Meier de la función de supervivencia está dado por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) \tag{2.8}$$

Donde:

n_j: número de individuos en riesgo al tiempo j

d_j: número de muertos al tiempo t_j.

2.2.5. Modelo de Cox

El modelo de riesgos proporcionales fue propuesto por Cox en 1972. Este modelo se basa en un grupo de variables o covariables independientes que influyen sobre la variable respuesta que es el tiempo de ocurrencia de un evento. El modelo asume independencia entre las observaciones de cada unidad. El modelo de riesgo especificado para el individuo i es:

$$\boldsymbol{\lambda_i}(t) = \lambda_0(t) e^{\mathbf{x_i}(t)\beta}, \tag{2.9}$$

Donde λ_0 es una función no negativa indeterminada de tiempo, llamada función de riesgo de línea base y β es un vector de coeficientes p x 1, quienes medirán el efecto de las covariables.

2.3. Eventos recurrentes

Kelly y Lim (2000) define un evento recurrente cuando un sujeto experimenta repetidas recurrencias del mismo tipo. Por otro lado Therneau y Grambsch (2000) los denomina eventos múltiples del mismo tipo o de diferente tipo, por ejemplo infecciones múltiples en pacientes con SIDA, infartos recurrentes en un estudio coronario, información recurrente en ensayos clínicos, rehospitalizaciones etc. La presente investigación, se utilizó con datos de eventos recurrentes del mismo tipo.

2.4. Conceptos teóricos relacionados a eventos recurrentes

- **Intervalo de riesgo**

El intervalo de riesgo en los datos de eventos recurrentes es considerado como un componente importante en el modelado; y, se define cuando un sujeto está en riesgo de tener un evento a lo largo de una escala de tiempo dado. (Kelly y Lim, 2000).

A continuación se ilustra tres formulaciones de intervalo de riesgo:

Cuadro 1: Estructura de los datos según formulación de tres tipos de intervalo de riesgo para un sujeto con cinco recurrencias

Individuo	Formulación					
	Brecha de tiempo (*Gap time*)		Proceso contador		Tiempo total (*Total time*)	
	T. Inicio	T. Fin	T. Inicio	T. Fin	T. Inicio	T. Fin
001	0	2	0	2	0	2
001	0	5	2	7	0	7
001	0	2	7	9	0	9
001	0	3	9	12	0	12
001	0	8	12	20	0	20

Fuente: Elaboración propia

En el Cuadro 1 se ejemplifica la forma como se estructuran los datos según tres tipos de intervalos de riesgo, para un sujeto con cinco recurrencias, las cuales servirán para modelar las diferentes metodologías expuestas en la presente investigación. La formulación tiempo de brecha (gap time) y proceso de conteo es usado para el modelo de Prentice, Williams y Peterson; la formulación proceso contador para el modelo Anderson-Gill y el tiempo total (Total time) para el modelo Wei, Lin y Weissfeld. Estos modelos se detallarán en el apartado 2.6. Una ilustración de estos tres tipos de intervalo de riesgo se presenta en la Figura 2.

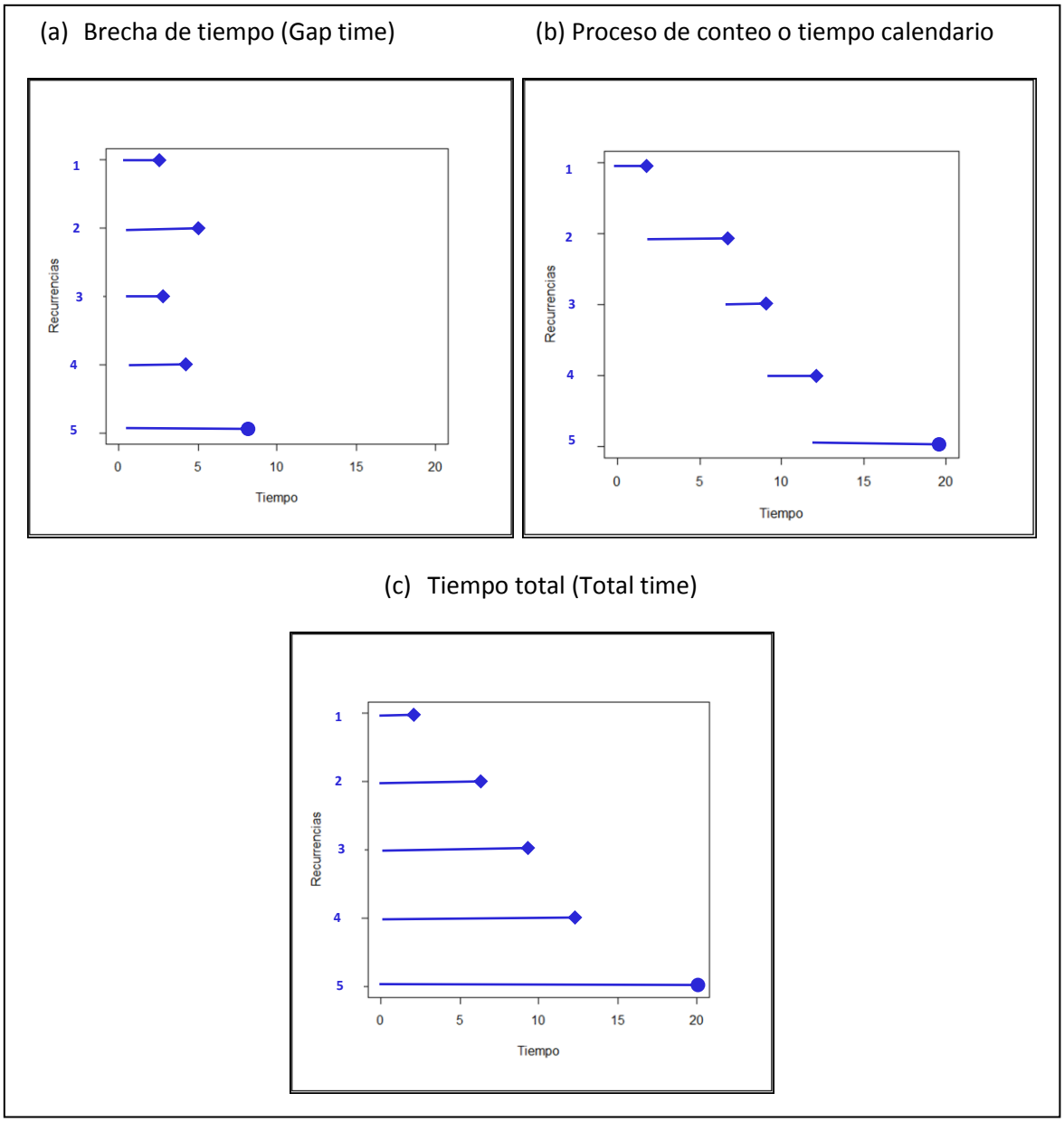


Figura 2: Ilustración de la formulación de tres tipos de intervalo de riesgo y representa a un sujeto con cinco recurrencias. El circulo (●) indica censura, y el rombo sólido (♦) indica ocurrencia de un evento.

La Figura 2 describe tres tipos de intervalo de riesgo para un sujeto con recurrencia de un evento. La figura 2(a) representa la formulación gap time que es el tiempo de un evento a priori, es decir se reanuda después de cada evento (Kelly y Lim, 2000) en donde el sujeto se encuentra en riesgo para su primer evento en el intervalo (0, 2], su segundo, tercero y

subsecuente evento durante los intervalos de tiempo (0,5], (0,2], (0,3], y (0,8] respectivamente.

El tiempo total o *total time* usualmente es el tiempo de inicio de tratamiento. En la figura 2(c) el sujeto está en riesgo en el primero, segundo hasta el quinto evento durante el intervalo (0,2], (0,7], (0,9], (0,12] y (0,20], respectivamente. El proceso de conteo usa similar escala de tiempo como tiempo total, sin embargo un sujeto podría tener entrada tardía o periodo de censura antes que el sujeto se encuentre en situación de riesgo, (Kelly y Lim, 2000). Se puede apreciar en la figura 2(b), el sujeto está en riesgo en los intervalos (0,2], (2,7], (7,9], (9,12] y (12,20], respectivamente.

Las formulaciones tiempo de brecha y proceso de conteo, tienen similar longitud de tiempo cuando están en riesgo. Los modelos de eventos recurrentes se construyen de acuerdo al tipo de intervalo de riesgo, y se clasifican como modelos marginales y condicionales.

- **Conjunto de riesgos (Risk Set)**

Es el *k*-ésimo conjunto de riesgos el cual contiene a los individuos quienes se encuentran en riesgo en el *k*-ésimo evento. El riesgo establecido en un momento dado depende de los individuos incluidos en el conjunto y cuando esos individuos están en riesgo, es decir, el intervalo de riesgo (Kelly y Lim, 2000).

2.5. Modelos de eventos recurrentes

2.5.1. Investigaciones relacionadas a modelos de datos de eventos recurrentes.

A inicios de los años 80, surge la idea que un evento puede ocurrir múltiples veces en el transcurso del seguimiento de un sujeto. Prentice et al. (1981) proponen dos modelos de eventos recurrentes, los cuales pueden ser considerados como extensiones del modelo estratificado de riesgos proporcionales de Cox (1972) con estrato definido por el evento recurrente. Por otro lado, Andersen et al. (1993), proponen un modelo de eventos recurrentes mediante la aproximación del proceso de conteo**.**

Wei et al. (1989), proponen modelos semi-paramétricos para analizar el tiempo multivariado de falla; es decir, modelar la distribución marginal de cada variable tiempo de falla con un modelo de riesgos proporcionales de Cox. Asimismo, los autores no consideran ninguna estructura particular de dependencia entre distintos tiempos de fallas en cada sujeto, por ello se asume que los tiempos no se encuentran correlacionados. Los parámetros de regresión son estimados mediante la maximización de la verosimilitud parcial de falla-específica.

Sullivan y Cai (1993), desarrollaron métodos de visualización gráfica y análisis de datos de tiempo de eventos múltiples, considerando dos problemas fundamentales: el primero, eventos de falla múltiples son del mismo tipo para cada individuo y el segundo problema fue el análisis del efecto de una covariable categórica dependiente del tiempo en la respuesta del tiempo de falla resultante.

Wang y Chang (1999), **s**e centran en la función de supervivencia marginal del tiempo entre dos eventos sucesivos, denominado la *función de supervivencia recurrente*. Un tema fundamental en la investigación de los autores son los tiempos de recurrencias de episodios diferentes que tienen la misma distribución. Usa los modelos de fragilidad para caracterizar la correlación de tiempos recurrentes del mismo sujeto, con la condición de censura independiente.

Peña et al. (2001), generalizan el estimador clásico de supervivencia de Kaplan Meier a eventos recurrentes en casos donde existe independencia entre los tiempos inter-ocurrencias. Los autores definieron dos procesos de conteo N e Y que permiten realizar estimaciones de su modelo.

Martines y Borges (2008), generalizan los modelos clásicos no paramétricos de análisis de supervivencia de eventos recurrentes, proponen estimadores de las funciones de supervivencia de manera gráfica, diseñando un programa en el software estadístico R. Estos estimadores son comparados con los propuestos por Peña et al. (2001).

Mazroui et al. (2012), propusieron un modelo de fragilidad conjunta para analizar las recurrencias y la muerte de forma simultánea. Dos distribuciones Gamma-Fragilidad tienen en cuenta tanto la dependencia entre recurrencias y la dependencia entre las recurrencias y

los tiempos de supervivencia. Además, estimaron parámetros independientes para la enfermedad de los tiempos recurrente de los eventos y los tiempos de supervivencia en el modelo de fragilidad conjunto para distinguir los efectos del tratamiento y factores pronósticos en estos dos tipos de eventos.

Androulakis et al. (2012), desarrollaron modelos de fragilidad con verosimilitud penalizada como un método a una función de verosimilitud general de datos organizados en grupos (cluster), lo que corresponde a una clase de modelos de fragilidad, que incluye el Modelo de Cox y el Modelo de Fragilidad Gamma como casos especiales. Asimismo, consideraron la formulación del modelo de fragilidad en el caso de las agrupaciones (o cluster), donde los sujetos en el mismo grupo comparten la misma fragilidad, que es una variable aleatoria positiva.

Rondeau et al. (2003), propusieron un Modelo de Fragilidad Gamma Compartida usando el método de máxima verosimilitud para penalizar la función de riesgo. Los resultados obtenidos por los autores los compara con el enfoque propuesto por Therneau y Grambsch (2000).

Rondeau et al. (2012), desarrollaron una nueva versión del paquete R llamado frailtypack, el cual permite ajustar modelos de Cox y cuatro tipos de modelos de fragilidad (compartida, jerarquizado, anidado, aditivos), además está adaptado para el análisis de eventos recurrentes.

2.6. Modelos de eventos recurrentes sin efecto aleatorio (sin fragilidad)

Se presenta tres modelos que pueden considerarse como extensiones del modelo de riesgos proporcionales de Cox. Estos modelos serán utilizados para identificar importantes factores pronósticos o de riesgo con la ayuda de software estadístico R. Los tres modelos son los siguientes: Prentice, Williams y Peterson; Andersen y Gill, y Wei, Lin y Weissfeld. Los tres modelos son extensiones de los modelos de riesgos proporcionales, y las funciones de verosimilitud de estos modelos se construyen de manera diferente, principalmente en el riesgo establecido en las observaciones no censuradas (Lee y Wenyu, 2003).

- **Notación General**

Sea N_i el número de eventos en el sujeto *i* y sea t_{ik} el verdadero tiempo total del k-ésimo evento en el i-ésimo sujeto, C_{ik} el tiempo de censura del k-ésimo evento en el i-ésimo sujeto y T_{ik} es tiempo correspondiente a la observación, donde $T_{ik} = \min(t_{ik}, C_{ik})$. Sea δ_{ik} la variable indicadora de falla en el k-ésimo evento y en el i-ésimo individuo; esto es, toma el valor de "1" si el evento es observado y "0" si es censurado. Sea λ_{ik} la función de riesgo base en el k-ésimo evento y en el i-ésimo sujeto, X_{ik} denota el vector de covariables para el i-ésimo sujeto con respecto al k-ésimo evento.

2.6.1. Modelo de incrementos independientes: Modelo Andersen – Gill (A-G)

Andersen y Gill (1982), asume independencia de eventos; es decir, que alguna ocurrencia de un evento no se ve afectada por eventos previos. Además, define el intervalo de riesgo usando la formulación de proceso de conteo (ver Cuadro 1), con un conjunto de riesgo no restringido y una base de riesgo común para todos los eventos.

Therneau y Grambsch (2002), indicó que este modelo es ideal para la situación de independencia mutua de las observaciones dentro de un sujeto. Esta suposición es equivalente a cada proceso de conteo individual que posee incrementos independientes, es decir, los números de eventos en intervalos de tiempo que no se superponen son independientes, dadas las covariables. Sin embargo, este supuesto puede relajarse incluyendo en el modelo una covariable dependiente de tiempo, de esta manera podría obtenerse la estructura de dependencia entre los tiempos de recurrencias. (Kelly y Lim, 2000).

Considere una secuencia de modelos indexados por $n = 1, 2, \ldots$, Andersen y Gill (1982), generalizan las observaciones posiblemente censuradas de los tiempos de vida de n individuos $N^{(n)} = (N_1^{(n)}, \ldots, N_n^{(n)})$, donde $N_i^{(n)}$ cuenta el número de eventos observados en la vida del i-ésimo individuo, $i = 1, \ldots, n$, sobre el intervalo $[0,1]$. Por lo tanto, el patrón de la muestra de $N_1^{(n)}, \ldots, N_n^{(n)}$ son funciones escalonadas, cero en el tiempo cero, con saltos de tamaño +1, sin procesos de dos componentes saltando al mismo tiempo.

La suposición básica es que para cada n, $N^{(n)}$ tiene proceso de intensidad aleatorio, también llamado función de riesgo $\lambda^{(n)} = (\lambda_1^{(n)}, ..., \lambda_n^{(n)})$ tal que:

$$\boldsymbol{\lambda}_i^n \left(t; \mathbf{x}_i\right) = Y_i^{(n)}(t)\lambda_0(t)\exp\{\beta_0^{'}\mathbf{x}_i^{(n)}(t)\} \tag{2.10}$$

Donde β_0 es un vector columna fija de *p* coeficientes, λ_0 es una función de riesgo subyacente, y $Y_i^{(n)}\left(t\right)$ es un proceso predecible que toma valores en {0,1}, donde 1 indica que el i-ésimo individuo está bajo observación. A diferencia del modelo clásico de Cox (1972), el individuo deja de estar en riesgo cuando $Y_i\left(t\right)$ toma el valor de cero, esto quiere decir que el evento ocurrió, sin embargo en el modelo propuesto por Andersen y Gill el valor que toma $Y_i\left(t\right)$ seguirá siendo uno si ocurre el evento ($N_i^{(n)}$ sólo saltos cuando $Y_i^{(n)} = 1$), dado que el individuo sigue en riesgo (Ver Figura 2). Finalmente $\mathbf{x}_i^{(n)} = (x_{i1}^{(n)}, ..., \mathrm{x}_{ip}^{(n)})'$ es un vector columna de *p* covariables para el i-ésimo individuo.

Entonces sea función de verosimilitud parcial extendido de Cox, 1972 está dado por:

$$L\left(\beta\right) = \prod_{i=1}^{n}\left(\frac{e^{\beta'\mathbf{x}_l(T_l)}}{\sum_{j\in\Re_l} Y_j(T_l)e^{\beta'\mathbf{x}_j(T_l)}}\right)^{\delta_l} \tag{2.11}$$

$$Log(L(\beta)) = \sum_{i=1}^{n}\int_0^t \beta' X_l(t)dN_i(\mathrm{t}) - \int_0^t \log\{\sum_{l=1}^{n} Y_i(\mathrm{t})\exp(\beta'\mathrm{X}_l(t))\}d\bar{N}(t)$$

Donde $\bar{N} = \sum_{i=1}^{n} N_i$. Entonces se estima el vector de parámetros $\hat{\beta}$ a partir de la solución de la ecuación que se obtiene de la derivada del logaritmo de la verosimilitud parcial en el tiempo t e igualando a cero y por métodos iterativos mediante el algoritmo de Newton Raphson. Asimismo se puede obtener el vector de derivadas y que tiene la forma:

$$U(\beta, \mathrm{t}) = \sum_{i=1}^{n}\int_0^t X_i(t)dN_i(t) - \int_0^t \frac{\sum_{i=1}^{n} Y_i(t)X_i(t)\exp(\beta' X_i(t))}{\sum_{i=1}^{n} Y_i(t)\exp(\beta' X_i(t))}d\bar{N}(t)$$

También se puede expresar como:

$$U(\beta_0, \mathrm{t}) = \sum_{i=1}^{n} \int_0^t X_i(t) dM_i(t) - \int_0^t \frac{\sum_{i=1}^{n} Y_i(t) X_i(t) \exp(\beta_0^{'} X_i(t))}{\sum_{i=1}^{n} Y_i(t) \exp(\beta_0^{'} X_i(t))} d\bar{M}(t)$$

Donde $\bar{M} = \sum_{i=1}^{n} M_i$ es un martingala local.

Los autores también probaron la normalidad asintótica asi como la propiedad de consistencia de $\hat{\beta}$. No se han mostrado las demostraciones de las propiedades por no formar parte de los objetivos del presente trabajo.

Therneau y Grambsch, 2000, mencionan en su libro que el modelo Andersen-Gill es eficiente y da la estimación más fiable del efecto general del tratamiento. El uso de la varianza robusta para el modelo AG no es necesario en teoría, pero en la práctica es la elección más sabia. Dos suposiciones básicas del modelo son que los eventos no cambian el sujeto (no hay cambio en el riesgo basal) y que es de interés una estimación general del efecto de la covariable.

2.6.2. Modelo Marginal: Modelo Wei, L, Lin Y. y Weissfeld (WLW)

Los modelos marginales asumen que los eventos dentro de un mismo sujeto son independientes y por lo tanto no está condicionado a la historia de los eventos o los eventos previos.

Este modelo utiliza la formulación como intervalo de riesgo el tiempo total o *total time* y asume un conjunto de riesgo semi-restringido y una base de riesgo de evento específico. El modelo utiliza la distribución marginal de cada variable de tiempo de falla con un modelo de riesgos proporcionales. No se considere ninguna estructura de dependencia entre los tiempos de falla de cada individuo. Los parámetros de regresión son estimados mediante la verosimilitud parcial de evento específico. Cada evento es considerado un estrato, y por lo tanto permite un riesgo subyacente separado para cada evento y por estrato por interacción de covariables. El indicador en riesgo para el *k-ésimo* evento, Y_{ik} *(t)* es uno hasta la

ocurrencia del evento *k*, a menos que, por supuesto, algún evento externo cause la censura y cuando ocurre, se convierte en cero (Ver figura N°3).

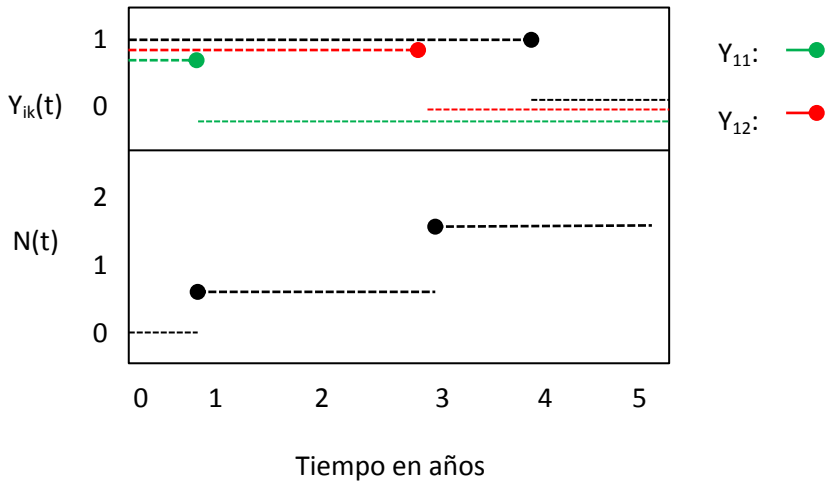


Figura 3: Ilustración del Indicador de riesgo (Y_{ik}) para el modelo WLW

Para el k-ésimo evento del i-ésimo individuo, la función de riesgo toma la forma siguiente:

$$\boldsymbol{\lambda}_{ik}\left(t;X_{ik}\right)=Y_{ik}\left(t\right)\lambda_{0k}(t)\exp(\beta_k{}'\mathbf{x}_i(t)),\ \ t\geq 0 \tag{2.12}$$

Donde $\lambda_{0k}(t)$ es una función de riesgo base no especificada y $\boldsymbol{\beta}_k=(\beta_{1k},...,\beta_{pk})$ son los parámetros de regresión de evento específico. La función de verosimilitud parcial para el k-ésimo evento específico está dado por:

$$L_k\left(\beta\right)=\prod_{i=1}^{n}\left(\frac{e^{\beta'\mathbf{x}_{ik}(T_{ik})}}{\sum_{l\in\Re_k(T_{ik})}e^{\beta'\mathbf{x}_{kl}(T_{ik})}}\right)^{\delta_{ik}} \tag{2.13}$$

Donde $\Re_k(\mathrm{t})=\left\{l:T_{kl}\geq t\right\}$ es el conjunto de sujetos que están en riesgo en el *k*-ésimo evento previo al tiempo *t*. La solución de la ecuación normal $\frac{\partial L_k(\beta)}{\partial\beta}=0$, es obtenida mediante la estimación de máxima verosimilitud parcial, de esta manera se obtiene los

estimadores $\hat{\boldsymbol{\beta}}_k$, siendo consistente para $\boldsymbol{\beta}_k$, siempre y cuando el modelo marginal (2.13) este correctamente especificada.

Propiedades Asintóticas de los estimadores de los parámetros

Wei et al. (1989), demostraron que para muestras grandes n, la distribución del vector aleatorio $(\hat{\beta}_1^{'},...,\hat{\beta}_k^{'})'$ puede ser aproximada por una distribución normal pk-dimensional con media $(\beta_1^{'},...,\beta_k^{'})'$ y con matriz de covarianzas que puede ser estimada. Para la k-ésima falla o evento, sea

$$N_{ik}(t) = I\{\mathrm{T}_{ik} \le \mathrm{t}, \delta_{ik} = 1\},$$
$$\mathrm{Y}_{ik}(\mathrm{t}) = I\{\mathrm{T}_{ik} \ge \mathrm{t}\},$$
$$M_{ik}(t) = N_{ik}(t) - \int_0^t Y_{ik}(u)\lambda_{ik}(u)du,$$

Donde $I\{\cdot\}$ es la función indicadora. Además de la expresión (2.13) se obtiene el logaritmo de la verosimilitud parcial dado por:

$$C_k(\beta;\mathrm{t}) = \sum_{i=1}^{n}\int_0^t \beta^{'} X_{ik}(u)dN_{ik}(u) - \int_0^t \log\left[\sum_{i=1}^{n} Y_{ik}(u)\exp\{\beta^{'} X_{ik}(u)\}\right] d\bar{N}_k(u),$$

Donde $\bar{N}_k(\mathrm{u}) = \sum_{i=1}^{n} N_{ik}(u)$. El vector de derivadas obtenido de $C_k(\beta;t)$ con respecto a β tiene la forma:

$$U_k(\beta;\mathrm{t}) = \sum_{i=1}^{n}\int_0^t X_{ik}(u)dN_{ik}(u) - \int_0^t \frac{\sum_{i=1}^{n} Y_{ik}(u)X_{ik}(u)\exp\{\beta^{'} X_{ik}(u)\}}{\sum_{i=1}^{n} Y_{ik}(u)\exp\{\beta^{'} X_{ik}(u)\}} d\bar{N}_k(u)$$

De esto se deduce que:

$$U_k(\beta_k;\mathrm{t}) = \sum_{i=1}^{n}\int_0^t X_{ik}(u)dM_{ik}(u) - \int_0^t \frac{\sum_{i=1}^{n} Y_{ik}(u)X_{ik}(u)\exp\{\beta_k^{'} X_{ik}(u)\}}{\sum_{i=1}^{n} Y_{ik}(u)\exp\{\beta_k^{'} X_{ik}(u)\}} d\bar{M}_k(u)$$

Donde $\bar{M}_k(\mathrm{u}) = \sum_{i=1}^{n} M_{ik}(u)$ y $U_k(\beta_k;\mathrm{t})$ es un martingala cuadrado local integrable en t con respecto al k-ésimo evento.

Por la serie de expansión de Taylor de $U_k(\beta_k;\infty)$ alrededor de β_k se obtiene $\hat{A}_k(\hat{\beta}_k^{*})$ que converge en probabilidad a una matriz no determinística y que puede ser estimada con

propiedades de consistencia. Entonces la matriz de covarianzas para muestras grandes n de $(\hat{\beta}_1^{'},...,\hat{\beta}_k^{'})'$ puede ser estimado por:

$$\hat{Q} = n^{-1}\begin{bmatrix} \hat{D}_{11}(\hat{\beta}_1,\beta_1)...\hat{D}_{1k}(\hat{\beta}_1,\hat{\beta}_k) \\ . \quad\quad\quad\quad\quad . \\ . \quad\quad\quad\quad\quad . \\ . \quad\quad\quad\quad\quad . \\ \hat{D}_{k1}(\hat{\beta}_k,\beta_1)...\hat{D}_{kk}(\hat{\beta}_k,\hat{\beta}_k) \end{bmatrix}$$

Donde $\hat{D}_{kl}(\hat{\beta}_k,\hat{\beta}_l) = \hat{\mathrm{A}}_k^{-1}(\hat{\beta}_k)\hat{B}_{kl}(\hat{\beta}_k,\hat{\beta}_l)\hat{A}_l^{-1}(\hat{\beta}_l)$ (mayor detalles de la demostración ver artículo de los autores).

El enfoque de WLW permite cambios en los efectos del modelo a lo largo del tiempo usando términos de interacción de estrato por covariable. El conjunto de datos para el modelo puede ser bastante grande. Sin embargo, la interpretación del modelo es menos clara, ya que la asunción de que un sujeto está "en riesgo" para el k-ésimo evento antes de que ocurra el evento k - 1 es un inconveniente. Además, el modelo puede violar gravemente la suposición de riesgos proporcionales (Therneau y Grambsch, 2000).

2.6.3. Modelo condicional: Modelo Prentice, Williams and Peterson (PWP)

Un sujeto no está en riesgo en el *k*-ésimo evento hasta que haya experimentado el evento *k-1st.* (Castañeda y Gerritse, 2010).

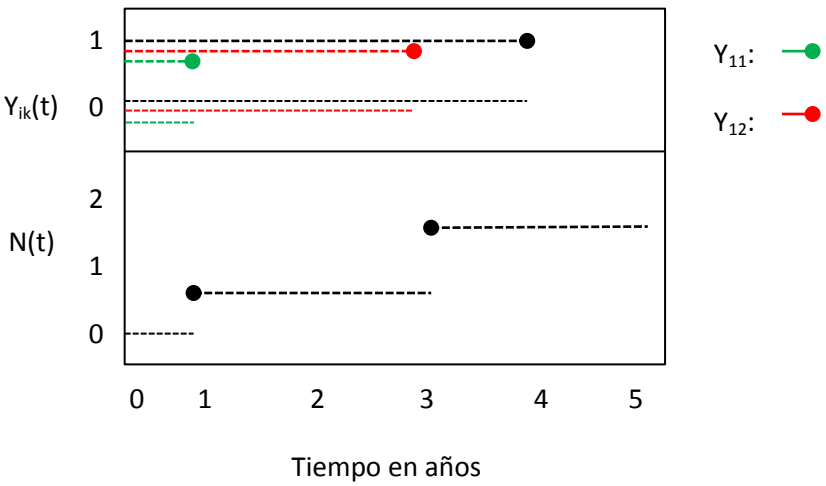


Figura 4: Ilustración del indicador de riesgo (Y_{ik}) para el modelo PWP

Este modelo usa estratos dependientes de tiempo el cual permite que el riesgo base varíe de evento a evento. La estratificación con el número previo de recurrencias controlará la dependencia entre eventos.

Cada evento es asignado a un estrato separado. La escala de tiempo que puede ser usado es proceso de conteo (Condicional I) y escala desde la entrada o tiempo de brecha (gap time) (Condicional II), para la presente investigación se usó la escala de proceso de conteo. La función de intensidad subyacente podría variar de evento a evento cuando el estrato es dependiente de tiempo y difiere del modelo de Andersen and Gill, porque asume que todos los eventos son idénticos. (Therneau y Grambsch, 2000).

La función de riesgo es una función para el *k*-ésimo sujeto *i*, el cual está dado por:

$$\lambda_{ik}\left(t;X_{ik}\right)=\lambda_{0k}(t)\exp(\beta_k{}'X_i(t)) \qquad (2.14)$$

Prentice et al. 1981, asumen que dado $\{N(t),X(t)\}$ los mecanismos de fallo o evento de los individuos en riesgo en el tiempo t actúan independientemente sobre $[t,t+dt]$, y bajo un supuesto de "censura independiente". Entonces se da la verosimilitud parcial considerando un modelo estratificado puede ser escrito como:

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{k} \left(\frac{\exp(\beta' X_{ik}(T_{ik}))}{\sum_{j=1}^{n} Y_{jk}(\mathrm{T}_{ik}) \exp(\beta' X_{ik}(T_{ik}))} \right)^{\delta_{ik}}$$

Para obtener los parámetros estimados se procede de forma similar a los modelos antes mencionados, $\frac{\partial \log(L(\beta))}{\partial \beta} = 0$, la derivada del logaritmo de la verosimilitud parcial y mediante el algoritmo de Newton-Raphson , se obtendrá los parámetros estimados.

Therneau y Grambsch, 2000, mencionan en su libro que el modelo condicional tiene la interpretación más natural, y es fácil de configurar en el software. Sin embargo, la existencia de factores de riesgo no modelados puede influir negativamente en los coeficientes. Los conjuntos de riesgo para los números de eventos posteriores también serán muy pequeños, haciendo que las estimaciones de riesgo por estrato sean inestables. Sin embargo a pesar de las imperfecciones de los modelos expuestos, los autores mencionan que estos proveen información importante.

2.6.4. Varianza Robusta

La varianza de los coeficientes en un modelo de Cox ordinario trata las observaciones como independientes. Sin embargo, cuando se tiene múltiples eventos esta suposición ya no se cumple debido a las correlaciones entre las observaciones para cada individuo, por lo que una corrección apropiada es usar la estimación agrupada de jackknife que deja un sujeto a la vez, en vez de una observación a la vez (Liptsitz, et al. 1990).

Therneau y Grambsch, 2000, proponen una forma de calcular los valores de jackknife para obtener la variancia de los coeficientes considerando valores individuales mediante iteración de Newton Rapson:

a. Obtener los coeficientes estimados $\hat{\beta}$ con todas las observaciones.

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + UI^{-1} \qquad (2.15)$$

Donde I^{-1} es la inversa de la matriz de información, y U es la matriz nxp de residuales score, entonces:

b. Retirar la observación i:

$$U_{(i)1xp}I^{-1} = D_{(i)} = \Delta\hat{\beta}_{(i)} = \hat{\beta}_{(i)}^{(k+1)} - \hat{\beta}_{(i)}^{(k)}$$

De tal manera que $D_{(i)}$ representa la i-ésima fila de la matriz $\mathbf{D}_{n\,x\,p}$ que es la matriz de cambio en $\hat{\boldsymbol{\beta}}$ si es removida la i-ésima observación.

$$\mathbf{D} = \begin{bmatrix} \hat{\beta}_{1(1)}^{(k)} - \hat{\beta}_{1(1)}^{(k-1)} & \hat{\beta}_{2(1)}^{(k)} - \hat{\beta}_{2(1)}^{(k-1)} & \cdots & \hat{\beta}_{p(1)}^{(k)} - \hat{\beta}_{p(1)}^{(k-1)} \\ \hat{\beta}_{1(2)}^{(k)} - \hat{\beta}_{1(2)}^{(k-1)} & \hat{\beta}_{2(2)}^{(k)} - \hat{\beta}_{2(2)}^{(k-1)} & \cdots & \hat{\beta}_{p(2)}^{(k)} - \hat{\beta}_{p(2)}^{(k-1)} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \hat{\beta}_{1(n)}^{(k)} - \hat{\beta}_{1(n)}^{(k-1)} & \hat{\beta}_{2(n)}^{(k)} - \hat{\beta}_{2(n)}^{(k-1)} & \cdots & \hat{\beta}_{p(n)}^{(k)} - \hat{\beta}_{p(n)}^{(k-1)} \end{bmatrix}_{nxp}$$

De tal manera que $\mathbf{D}'\mathbf{D}$ es denominado estimador de varianza sándwich siendo esta la varianza robusta de observaciones individuales independientes.

La varianza sándwich puede ser escrita como:

$$\mathbf{D}'\mathbf{D} = \mathbf{I}^{-1}\mathbf{U}'\mathbf{U}\mathbf{I}^{-1}$$

Con grupos correlacionados el estimador sándwich estaría dado por $\tilde{\mathbf{D}}'\tilde{\mathbf{D}}$, donde $\tilde{\mathbf{D}}_{mxp} = \mathbf{B}_{mxn}\mathbf{D}_{nxp}$, y, $\mathbf{B}$ es una matriz de ceros y unos que suma la propia fila, y que podría funcionar como una variable indicadora, tal que $\tilde{\mathbf{D}}_{m\,x\,p} = \mathbf{B}\mathbf{U}\mathbf{I}^{-1}$, entonces el cálculo de la varianza robusta para datos correlacionada, está dada por lo siguiente:

$$\tilde{\mathbf{D}}'\tilde{\mathbf{D}} = \mathbf{I}^{-1}\mathbf{U}'\mathbf{B}'\mathbf{B}\mathbf{U}\mathbf{I}^{-1} \qquad (2.16)$$

2.6.5. Estimación e Inferencia

El estimador de máxima de verosimilitud parcial se encuentra resolviendo la ecuación $U\left(\hat{\beta}\right)=0$ mediante el algoritmo de Newton-Raphson, con procesos iterativos igual a (2.15), donde la estimación inicial $\beta=0$, $\mathbf{I}^{-1}$ es la inversa de la matriz de información, el cual es utilizada como la varianza de $\boldsymbol{\beta}$.

Prueba Global:

$H_0:\beta=\beta^{(0)}$

Ratio de Verosimilitud

La prueba de ratio de verosimilitud se obtiene dos veces la diferencia en la log verosimilitud parcial en la estimación inicial y final de $\hat{\beta}$. Está dado por:

$$2\left(l\left(\hat{\beta}\right)-l\left(\beta^{(0)}\right)\right) \tag{2.17}$$

Prueba Robusta de Wald

Debido a la recurrencia del evento de interés, los tiempos por sujeto se encuentran correlacionados, y la estimación del sandwich $\tilde{\mathbf{D}}'\tilde{\mathbf{D}}$ será frecuentemente mayor que la variante $\mathbf{I}^{-1}$ basada en el modelo, por lo que Therneau y Grambsch, consideran a las pruebas usuales como no conservadoras y utiliza la prueba robusta de Wald, el cual está basado en la estimación de sándwich dado:

$$\hat{\boldsymbol{\beta}}'[\tilde{\mathbf{D}}'\tilde{\mathbf{D}}]^{-1}\hat{\boldsymbol{\beta}} \tag{2.18}$$

Prueba Robusta Score

En esta prueba también es posible ajustar cuando existe correlación en los tiempos, y el estadístico para la prueba de score está basado en la primera iteración para estimar $\hat{\beta}$ con

el algoritmo de Newton-Raphson, y viene dado por $S = \left[\mathbf{1}^{'}\mathbf{U}\right]\mathbf{I}^{-1}\left[\mathbf{U}^{'}\mathbf{1}\right]$. Sin embargo, al insertar la inversa del estimador sándwich de varianza basado también en la iteración inicial de β , y recordando que $\mathbf{I} = \mathbf{U}^{'}\mathbf{U}$, por lo que si se considera la variable indicadora como corrección $\tilde{\mathbf{I}} = \mathbf{U}^{'}\mathbf{B}^{'}\mathbf{B}\mathbf{U}$, se obtiene el estadístico de la prueba de score:

$$S_r = \left[\mathbf{1}^{'}\mathbf{U}\right]\left[\mathbf{U}^{'}\mathbf{B}^{'}\mathbf{B}\mathbf{U}\right]^{-1}\left[\mathbf{U}^{'}\mathbf{1}\right] \tag{2.19}$$

Donde la primera iteración es $\hat{\beta}^{(1)} = \hat{\beta}^{(0)} + \mathbf{U}\mathbf{I}^{-1}$, además U y I^{-1} se calculan de $\hat{\beta}^{(0)}$.

La distribución de cada uno de los estadísticos para probar la hipótesis nula $H_0 : \beta = \beta^{(0)}$ es una chi-cuadrado con p grados de libertad.

Intervalo de confianza

Un intervalo (1-α)x100 confianza para el riesgo, basado en la varianza robusta, está dado por:

$$IC\left(\lambda(t)\right) = \left[\exp\left\{\hat{\beta} - Z_{\left(1-\frac{\alpha}{2}\right)}EE\right\}, \exp\left\{\hat{\beta} + Z_{\left(1-\frac{\alpha}{2}\right)}EE\right\}\right] \tag{2.20}$$

Donde EE indica el error estándar robusto. La librería survival permite obtener los estimadores de los intervalos de confianza para el riesgo.

2.6.6. Diagnóstico evaluación de influencia

Therneau y Grambsch, (2000), realiza la descripción del análisis residual y medida de influencia usando el valor de *Jackknife* agrupado que evalúa cada punto en el ajuste del modelo, definido como:

$$\mathbf{j}_i = \hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}} \tag{2.21}$$

Donde $\hat{\boldsymbol{\beta}}_{(i)}$ es el resultado de un ajuste, el cual generalizando para eventos múltiples incluiría todos los individuos excepto el individuo i. Entonces el residuo *Jackknife*_se

puede calcular usando el mismo esquema iterativo del algoritmo (2.15), de tal manera que la matriz de valores influyentes Jackknife estaría conformado por:

$$\mathbf{J} = \begin{bmatrix} \hat{\beta}_{1(1)} - \hat{\beta} & \hat{\beta}_{2(1)} - \hat{\beta} & \dots & \hat{\beta}_{p(1)} - \hat{\beta} \\ \hat{\beta}_{1(2)} - \hat{\beta} & \hat{\beta}_{2(2)} - \hat{\beta} & \dots & \hat{\beta}_{p(2)} - \hat{\beta} \\ \cdot & & \cdot & \\ \cdot & & \cdot & \\ \cdot & & \cdot & \\ \hat{\beta}_{1(n)} - \hat{\beta} & \hat{\beta}_{2(n)} - \hat{\beta} & \dots & \hat{\beta}_{p(n)} - \hat{\beta} \end{bmatrix}_{nxp}$$

Para un modelo lineal, el residuo de jackknife se puede calcular de múltiples maneras, todas dando el mismo resultado. Sin embargo debido a la cantidad de cálculos computacionales lo más simple de proceder a la iteración de Newton Raphson para el modelo de Cox $\mathbf{\Delta\beta = 1'\left(UI^{-1}\right) = 1'D}$, donde la matriz **D** es llamada matriz de residuales *dfbeta*, y U también puede ser obtenido como vector de score:

$$\mathbf{U} = \begin{bmatrix} \frac{\partial l}{\partial \beta_{1(1)}} & \frac{\partial l}{\partial \beta_{2(1)}} & \dots & \frac{\partial l}{\partial \beta_{p(1)}} \\ \frac{\partial l}{\partial \beta_{1(2)}} & \frac{\partial l}{\partial \beta_{2(2)}} & \dots & \frac{\partial l}{\partial \beta_{p(2)}} \\ \cdot & & \cdot & \\ \cdot & & \cdot & \\ \cdot & & \cdot & \\ \frac{\partial l}{\partial \beta_{1(n)}} & \frac{\partial l}{\partial \beta_{2(n)}} & \dots & \frac{\partial l}{\partial \beta_{p(n)}} \end{bmatrix}_{nxp}$$

2.6.7. Evaluación de la forma funcional de covariables

Cuando los datos presentan censura a la derecha y todas las covariables son fijadas al inicio del estudio, es decir, no son dependientes del tiempo, los residuales martingalas se utilizan para verificar la forma funcional de las covariables o presencia de valores atípicos. (Colosimo y Ruiz, 2006).

El residual martingala está dado por:

$$\begin{aligned}\hat{M}_i(t) &= N_i(t) - \hat{E}_i(t) \\ &= N_i(t) - \int_0^t Y_i(s) e^{X_i(s)\hat{\beta}} d\hat{\Lambda}_0(s)\end{aligned} \tag{2.22}$$

El vector $\hat{\beta}$ es el estimador máximo verosímil parcial, y $\hat{\Lambda}_0$ es el estimador de riesgo base acumulada. El residual martigala es la diferencia O-E entre el número observado de eventos para un individuo y su número esperado dado el modelo ajustado en el tiempo de seguimiento. Si un individuo está representado en una tabla por múltiples intervalos, el residual martingala para el i-ésimo sujeto es la suma de los residuales de sus observaciones. (Therneau y Grambsch, 2000)

2.6.8. Prueba de riesgos proporcionales

Therneau y Grambsch, (2000), proponen un estadístico para el supuesto de riesgos proporcionales el cual utiliza los residuales escalados de schoenfeld.

Se puede escribir $\beta(t)$ como una regresión en $g(t)$ dado en la siguiente expresión:

$\beta_j(t) = \beta_j + \theta_j(g_j(t) - \bar{g}_j)$, $j = 1,..., p$, $\bar{g}_j$ es la media de las funciones específicas $g_j(t_k)$, es decir la j-ésima variable, k-ésimo evento.

La varianza denotada como $\hat{V}(\beta, t)$, para mayor conjunto de datos es poco estable y cambia lentamente a medida que aparecen los últimos eventos. Entonces sabiendo que $\sum_k \hat{V}_k = I(\hat{\beta})$, sugiere el uso del valor promedio $\bar{V} = I / d$.

Sea la matriz d x p de residuales schoenfeld escalado $S^* = dSI^{-1}$, bajo el supuesto de varianza constante, se asume que $g_j(t) \equiv g(t)$, es decir, similar test está siendo usado para todas las covariables en un modelo dado. Entonces el test global de riesgos proporcionales sobre todas las p covariables está dado por:

$$T = \frac{(g - \bar{g})' S^* I S^* ' (g - \bar{g})}{d \sum (g_k - \bar{g})^2} \tag{2.23}$$

Sea $I^{jk} = I_{jk}^{-1}$, el *(j,k)* elemento de I^{-1}. Entonces la prueba de proporcionalidad para la j-ésima covariable está dado por:

$$T_j = \frac{\{\sum_k (g_k - \bar{g}) s_{kj}^*\}^2}{dI^{jj} \sum_k (g_k - \bar{g})^2} \tag{2.24}$$

(2.23) y (2.24) tienen una distribución X_p^2 con p grados de libertad. Extensión del estadístico de riesgos proporcionales pueden obtenerse para eventos múltiples con el comando cox.zph de la librería Survival (Colosimo y Ruiz, 2005).

2.7. Modelo de fragilidad para eventos recurrentes

El riesgo de algún evento en el tiempo depende en parte de una variable aleatoria no observable, que se supone actúa multiplicativamente sobre el riesgo, de manera que un valor grande de la variable, aumenta el riesgo a lo largo de todo el intervalo de tiempo (Nielsen et. *al*, 1992). Entonces, varios eventos pueden ocurrir a similar paciente, y la fragilidad individual podrían influir en la recurrencia del evento de interés (Rondeau, V. y Gonzalez J. 2005).

- **Modelo de fragilidad compartida**

El modelo de fragilidad compartida extiende el modelo de riesgos proporcionales de Cox, teniendo en cuenta la heterogeneidad no observable entre individuos (Rondeau, 2010).

Dado que las fragilidades son vistas como covariables no observables, un algoritmo bastante usado es el algoritmo EM, el cual para los modelos de fragilidad compartida tiene la desventaja de realizar las estimaciones de los coeficientes de manera lenta. Es por ello que surgen modelos penalizados como una alternativa de aproximación y los términos de fragilidad son tratados como coeficiente de regresión adicional, el cual están limitados por una función de penalización que agregan al log-verosimilitud. (Therneau *et al.* (2012). Sin embargo, Rondeau y Gonzalez (2005), mencionan que esos métodos presentan inconvenientes debido a la lenta convergencia para estimar los parámetros, asimismo no estima la varianza de la fragilidad y el método no puede ser usado para estimar la función de riesgo, el cual es de gran importancia para la interpretación en el área de epidemiología.

Por lo tanto, los autores han usado un método para la estimación no paramétrica de la función de riesgo usando un estimador continuo basado en la verosimilitud penalizada completa, la solución usada es aproximación con splines. Asimismo considera que el tiempo de los eventos de similar sujeto podría estar fuertemente correlacionados.

La escala de tiempo usada para el modelo de fragilidad es el tiempo calendario o proceso de conteo.

- **Notación General:**

Sea el sujeto i $(i=1,...,N)$, $(j=1,...,n_i)$ la recurrencia j, t_{ij} es el tiempo de *j*-ésima recurrencia del *i*-ésimo individuo; C_i tiempo de censura; $T_{ij}=\min(t_{ij},C_i)$ correspondiente al tiempo de seguimiento, y $\delta_{ij}=I_{(T_{ij}=t_{ij})}$, donde $I_{(\ )}$ denota la función indicadora. Sea $X_{ij}=(X_{1ij},...,X_{pij})$ vector de covariable fijas o dependiente de tiempo por individuo i. Sea ω_i el efecto aleatorio con distribución gamma.

- **Modelo:**

El riesgo en el tiempo t_{ij} para un individuo con fragilidad compartida, ω_i está dado por:

$$\lambda_{ij}(t_{ij})=\lambda_o(t_{ij})\omega_i\exp(\beta' X_{ij}) \tag{2.25}$$

Donde λ_0 es la función de riesgo base, ω_i es el efecto aleatorio común que toma en cuenta la dependencia entre eventos sucesivos dentro de un mismo paciente, y refleja mejor el verdadero curso clínico de la enfermedad en su población heterogénea (Rondeau, V. 2010). Las fragilidades ω_i se asumen independiente e idénticamente distribuida como una Gamma con media 1 y varianza desconocida θ, es decir $\omega_i \sim \Gamma\left(\frac{1}{\theta},\frac{1}{\theta}\right)$. Un valor de varianza alta indica alta correlación entre los tiempos de los eventos por individuo (Rondeau et al. (2012). La función de densidad de probabilidad para la fragilidad está dado por:

$$g(\omega)=\frac{\omega^{(1/\theta-1)}\exp\{-\omega/\theta\}}{\Gamma(1/\theta)\theta^{1/\theta}} \tag{2.26}$$

Cada sujeto presenta distintos periodos de riesgo de tal manera que n_i periodos por sujeto i, entonces cada sujeto será representado por una terna $(Y_{i11}, Y_{i12}, \delta_{i1}) \ldots (Y_{in_i1}, Y_{in_i2}, \delta_{in_i})$, donde la *j*-ésima terna, Y_{ij1} es el inicio del j-ésimo periodo de riesgo, Y_{ij2} es el término del *j*-ésimo periodo de riesgo y δ_{ij} es el indicador de censura.

A continuación se presenta la función log-verosimilitud completa:

$$l(\lambda_o(.), \beta, \theta) = \sum_{i=1}^{G} \left\{ \sum_{j=1}^{n_i} \delta_{ij} \left\{ \boldsymbol{\beta}' X_{ij}(Y_{ij2}) + \ln\left(\lambda_o\left(Y_{ij2}\right)\right) \right\} - \left(\frac{1}{\theta} + m_i\right) \right.$$

$$\text{x } \ln\left[1 + \theta \sum_{j=1}^{n_i} \left(\Lambda_0\left(Y_{ij2}\right) - \Lambda_0\left(Y_{ij1}\right)\right) \exp(\boldsymbol{\beta}' X_{ij}\left(Y_{ij2}\right)\right] + I_{\{m_i \neq 0\}} \sum_{k=1}^{m_i} \left[\ln\left(1 + \theta\left(m_i - k\right)\right)\right] \Bigg\} \quad (2.27)$$

Donde $\Lambda_0(.)$ es la función de riesgo acumulada y $m_i = \sum_{j=1}^{j_i} I_{\{\delta_{ij}=1\}}$ es el número de eventos observados en el *i*-ésimo individuo.

Rondeau et al. (2003), propone el método de estimación máxima log-verosimilitud penalizada, que consiste en suavizar la función de riesgo base penalizándola por medio de un término que toma grandes valores para funciones rugosas, de esta forma penaliza la función de riesgo a diferencia de Therneau y Grambsch (2002) que penaliza las fragilidades. Entonces la función log-verosimilitud penalizada está dado por:

$$pl(\lambda_0(.), \beta, \theta) = l(\lambda_0(.), \beta, \theta) - k \int_0^{\infty} \lambda^{"2}(t)\, dt \quad (2.28)$$

Donde $l(\lambda_0(.), \beta, \theta)$ es la log-verosimilitud completa dado en (2.27), y $\kappa \geq 0$ es un parámetro suavizado que controla el equilibrio entre el ajuste de datos y la suavidad de las funciones. Asimismo este parámetro puede ser fijado por el investigador o puede ser estimado por el método de maximización de un criterio de verosimilitud de validación cruzada para el modelo de Cox (Joly et al. 1998), el cual estima un valor de parámetro suavizado minimizando la siguiente función:

$$\bar{V}(k) = \frac{1}{n} \left\{ tr\left(\hat{H}_{pl}^{-1} \hat{H}_l - l\left(\hat{\Phi}_k\right)\right) \right\} \quad (2.29)$$

Donde $\hat{\Phi}_k$ es el estimador máximo verosímil penalizado, $\hat{\mathbf{H}}_{\mathbf{l}}$ es la matriz Hessiana convergente de la log-verosimilitud, $\hat{\mathbf{H}}_{\mathbf{pl}}$ es la matriz Hessiana convergente de la log-verosimilitud penalizada. Para minimizar $\bar{V}(k)$, se calcula varios valores alejados de k, evitando el mínimo local. (Rondeau, et al., 2012).

El estimador "sándwich" $\hat{\mathbf{H}}_{\mathbf{pl}}^{-1}\mathbf{I}\hat{\mathbf{H}}_{\mathbf{pl}}^{-1}$ puede ser usado como el estimador de la varianza de los parámetros del modelo dado en (2.28), donde I es la matriz de información (Rondeau, V. y Gonzalez J. 2005).

- **Maximización de la verosimilitud penalizada**

Para la estimación de los parámetros de la ecuación 2.28 se usa el algoritmo robusto Marquardt (Marquardt, 1963), que es la combinación de dos algoritmos; Newton Raphson y un algoritmo de descenso o pendiente más empinado. Este algoritmo asegura tener funciones positivas de riesgo en todas sus etapas, la varianza de las fragilidades y los coeficientes splines deben ser positivas, por lo tanto se asegura positividad con el uso de la transformación cuadrada (Rondeau et al. 2012). El paso de Newton implica una línea de búsqueda, y si el nuevo punto no es mejor, entonces ocurre una reducción. El paso de descenso o pendiente empinado envuelve una línea de búsqueda completa y se intenta sólo si el paso de Newton ha fracasado, debido generalmente a una dificultad de encontrar la inversa de la log-verosimilitud Hessiana, asimismo, en otros casos, la iteración de pendiente más pronunciada se utiliza con frecuencia porque la Hessiana puede ser singular y la convergencia es lenta (Joly *et al.* 1998).

El vector de parámetros se actualiza hasta la convergencia usando la siguiente iteración:

$$\Phi^{(r+1)} = \Phi^{(r)} - \delta\left(H^{(r)}\right)^{-1}\Delta\left(L\left(\Phi^{(r)}\right)\right) \qquad (2.30)$$

Donde:

δ: toma el valor de 1 por defecto(en el programa), sin embargo puede ser modificado para asegurar que la verosimilitud sea mejorada en cada iteración.

$\tilde{H}$: es una matriz Hessiana diagonalmente influida para asegurar una definición positiva.

$\Delta\left(L\left(\Phi^{(r)}\right)\right)$: corresponde a la gradiente de log-verosimilitud penalizado en la r-ésima iteración.

Las iteraciones finalizan cuando la diferencia entre dos log-verosimilitud consecutivas son pequeñas, los coeficientes son estables y el gradiente es suficientemente pequeño ($<10^{-4}$) (Joly et al. 1998, Rondeau et al. 2012). Además la primera y la segunda derivada son calculados usando el método de diferencias finitas. Asimismo los errores estándar de las estimaciones se obtienen de la inversa de la matriz Hessiana.

- **Estimación de la función de riesgo base**

Rondeau et al. (2005), mencionan que una función spline es definida como una secuencia de nudos creciente con coeficientes $\eta = (\eta_1, ..., \eta_m)'$. Este enfoque de Spline puede ser usado para estimar la función de riesgo base $\lambda_0(.)$ con k nodos: $\tilde{\lambda}_0(.) = \sum_{i=1}^{m} \eta_i M_i(.)$. En la propuesta de los autores utilizan M-spline cubico, es decir funciones polinomiales de 3er orden, que es una combinación lineal para aproximar una función en intervalos Ramsay J. (1988). La suma de funciones polinomiales es usado para aproximar la segunda derivada de la función de riesgo base $\lambda_0^{''}(.)$. Esta aproximación permite formas flexibles de la función de riesgo, reduciendo al mismo tiempo el número de parámetros. A mayor número de nodos que se use, más cerca es la aproximación a la función de riesgo verdadero.

Rondeau et al. (2012), mostraron una aproximación de intervalos de confianza al 95% para $\lambda_0(.)$ y está dado por:

$$\tilde{\lambda}_o(t) \pm 1.96\sqrt{M(t)^T I_{\hat{\eta}}^{-1} M(t)} \qquad (2.31)$$

Donde $I_{\hat{\eta}} = \frac{\partial^2 pl(\hat{\eta})}{\partial \eta^2}$ y $M(t) = (M_1(t), ..., \mathrm{M}_m(t))$ es el vector M-spline.

2.8. Aplicaciones y Casos prácticos de modelos de eventos recurrentes.

Barceló (2002), estudió los modelos de Andersen - Gill (incrementos independientes), modelos marginales, modelos condicionales, modelo de Fragilidad de Andersen-Gill usando un algoritmo de estimación modificado EMB y modelo de Fragilidad de Cox penalizado maximizando la función de verosimilitud parcial penalizada, todos ellos en el análisis de supervivencia multivariante. Para la ilustración de estos modelos el autor usó una base de datos de una cohorte prospectiva de pacientes admitidos en un mínimo de 24 horas en un hospital de Girona, Barcelona-España, y en un período comprendido entre el 15 de marzo al 15 de junio de 1999. La variable en estudio consiste en el tiempo transcurrido entre el ingreso en la UCI hasta la aparición de la infección, y que además presentaron eventos sucesivos de infección por paciente, los cuales también fueron registrados. Entre sus resultados que obtuvo el autor, encontró que el algoritmo EMB de estimación del modelo de Andersen-Gill de fragilidad presenta mejores resultado, esto debido a que presenta intervalos de confianza más reducidos, mejor ajuste con el criterio AIC y capturar la dependencia serial y heterogeneidad individual.

Baye, F. 2011, realizó una evaluación de los modelos marginales, condicionales y modelo de fragilidad, con la finalidad de comparar los diferentes métodos y predecir futuros eventos recurrentes en pacientes con cáncer a la vejiga. En la investigación se realizó el seguimiento a 615 pacientes con el cáncer mencionado durante los años 1974 a 2011. Los autores estudiaron los modelo de Andersen-Gill, Condicional I, Condicional II (PWP), Marginal (WLW) y el modelo de Fragilidad Compartida Gamma, esta última usó el algoritmo EM (Esperanza-Maximización) para el problema de estimación de parámetros. El autor analizó desde la primera hasta la cuarta recurrencia con cada modelo y entre sus resultado encontrados obtuvo que el género, multiplicidad del tumor, estadío y número de recurrencias previas, fueron los factores de riesgo asociadas a la recurrencia de cáncer de vejiga. Asimismo, el modelo de Andersen-Gill y Fragilidad Gamma Compartida con escala de tiempo de proceso de conteo presentaron mejor aproximación debido a su habilidad predictiva y discriminación.

III. MATERIAL Y MÉTODOS

3.1. Materiales

Para la presente investigación se utilizó los siguientes materiales:

- Una computadora Intel core i5
- Una impresora HP Laser Jet P1102w
- Disco externo Toshiba 1 tera
- Tres millares de papel bond A4
- Programas estadísticos:
- Software R versión 3.1.3, dentro del cual se usaron las siguientes librerías: Survival y frailtypack.

3.2. Metodología de la investigación

3.2.1. Tipo de la investigación

La presente investigación es de tipo descriptivo y correlacional, ya que se busca conocer las relaciones encontradas de las covariables con el tiempo de recurrencia de cáncer de mama observados en pacientes con dicho diagnóstico.

3.2.2. Diseño de la investigación

El diseño de investigación es no experimental, longitudinal de tipo cohorte, dado que estudia la ocurrencia de un evento específico de un grupo de individuos, los cuales se encuentran libres del evento (o enfermedad) al inicio del estudio, y se realiza el seguimiento a través del tiempo hasta la ocurrencia del evento (Lazcano, E. *et.* al. 2000, Hernández, M. 2009).

La presente investigación se hace un seguimiento a los pacientes a través del tiempo mediante las historias clínicas del cual se extrajo una muestra desde el año 2008 al 2016.

3.3. Población y Muestra

La población en estudio está conformada por una cohorte de todos los pacientes con diagnóstico de cáncer de mama, identificados en las historias clínicas del Instituto de Enfermedades Neoplásicas (INEN). Cabe mencionar que los estudios de cohorte es el seguimiento de individuos en el tiempo y que presentan una característica en común.

La muestra utilizada es el registro de historias clínicas de todos los pacientes con recurrencia de cáncer de mama durante el periodo el enero del 2008 hasta Agosto del 2016 del Instituto de Enfermedades Neoplásicas (INEN).

3.4. Descripción de los Datos

Para ilustrar cada uno de los modelos expuestos en la presente investigación se utilizó una base de datos de 68 pacientes mujeres mayores de 26 años de edad con diagnóstico de cáncer de mama, quienes fueron sometidas a cirugía mastectomía en el Instituto Nacional de Enfermedades Neoplásica (INEN) en la ciudad de Lima-Perú. Las características de los pacientes fueron extraídas de las historias clínicas, dentro de la cual algunas variables no fueron consideradas, debido a la carencia de información en los documentos antes mencionados.

3.5. Identificación de las variables

Los datos recopilados es una cohorte de historias clínicas que muestran la recurrencia de cáncer de mama en pacientes sometidos a cirugía mastectomía, durante un periodo de tiempo de seguimiento de 2008 al 2016 según historia clínica de paciente. En ellas se analizaron variables relacionadas al tiempo hasta recurrencia de cáncer o condición actual de la paciente (es decir tiempo de censura), características clínicas o covariables: Edad que fue diagnosticado el cáncer, Edad de la primera menstruación (Menarquia) de la paciente, si la paciente estaba o no en estado de Menopausia cuando fue diagnosticada con la

enfermedad, Paridad: multípara o nulípara, tipo histológico (TipoHist), grado del tumor (Grado), tamaño del tumor al inicio del estudio (TamañoTum), CompaGang.: compromiso ganglionar.

En el siguiente cuadro, se describen las variables que se han incluido en el estudio.

Cuadro 2: Descripción de las covariables incluidas en la investigación.

Variables	Descripción	Categoría de la variable
Edad	Edad del paciente	-
Menarquia	Edad de su primera menstruación	-
Menopausia	Paciente en estado de menopausia al diagnóstico	Si - No
Paridad	Multípara o Nulípara	Multípara - Nulípara
TipoHist	Tipo histológico	Ductual – Lobulillar-Otro
Grado	Grado del tumor	I, II y III
TamañoTum	Tamaño del tumor	-
CompGang	Compromiso ganglionar si es en la axila (ganglios linfáticos).	Si - No

Fuente: Elaboración propia

La Sociedad Americana de Cáncer, afirma que el cáncer de mama inicia cuando algunas células en el pecho comienzan a crecer fuera de control. Estas células suelen formar un tumor que a menudo se puede ver en una radiografía o se siente como un bulto. Por ello, para que en la presente investigación se comprenda mejor cada una de las variables estudiadas relacionadas al cáncer de mama se menciona una breve definición de cada una.

Edad: Se consideró la edad (en años) del paciente al inicio del estudio; es decir, después de someterse a una cirugía de mastectomía.

Menarquia: es la edad de inicio de la menstruación, que es el derramamiento periódico del revestimiento uterino en mujeres en edad reproductiva (Solages, M. 2013).

Menopausia: es el cese definitivo de la menstruación resultante de la secreción reducida de la hormona ovárica que ocurre naturalmente o es inducida por la cirugía, la quimioterapia, o la radiación (Mohammed, E. 2012). Para la presente investigación se consideró mujeres con menopausia natural, y como variable indicadora si-no.

Paridad: Según Oxford Medical Dictionary, indica el número de embarazos que tuvo una mujer y que ha dado lugar al nacimiento de un bebé capaz de sobrevivir. Asimismo, una multípara, como una mujer que ha dado a luz a un niño vivo después de cada uno de al menos dos embarazos y nulípara que no ha dado lugar a ningún nacimiento.

Tipo Histológico: La Organización mundial de la salud clasifica los tipos de carcinoma o neoplasia maligna epitelial: carcinoma ductual (se ubica en los conductos mamarios, donde el cáncer sigue un proceso de filtración hasta el tejido adiposo), carcinoma lobulillar (sigue el mismo proceso de filtración que el carcinoma ductual) y otros menos frecuentes.

Grado de Tumor: Según el Instituto de Nacional de Cáncer, es una indicación de la rapidez con la que probablemente se extenderá el cáncer y crecerá el tumor.

Tamaño del tumor: el tamaño indica a la extensión transversal del tumor en su punto más ancho, medidos en milímetros.

Compromiso ganglionar: si el cáncer de mama afectó o no los ganglios linfáticos.

En la presente investigación se usó tres tipos o formulaciones de intervalo de riesgo: tiempo de brecha (gap time), tiempo total y proceso de conteo, descritos en el capítulo anterior (Ver Cuadro 1). Asimismo, se consideró la variable enum que indica el número de recurrencia, el cual estratificará los eventos según intervalo de riesgo, la variable evento donde “1” indica si presenta recurrencia de la enfermedad de cáncer, y “0” indica censura, y por último la escala de tiempo es en días.

3.6. Metodología Aplicada

Se realizó la comparación de modelos de eventos recurrentes; incrementos independientes A-G, modelos marginales (WLW) y condicionales (PWP), así como el modelo de Fragilidad Compartida Gamma mencionados y descritos en el capítulo anterior.

A continuación se detalla los pasos realizados para el análisis de los datos:

1. Análisis exploratorio y descriptivo de las variables en estudio
2. Determinar los modelos ajustados de eventos recurrentes.

Modelo Anderson-Gill (A-G) dado en la ecuación (2.10)

$$\boldsymbol{\lambda}_{\mathbf{i}}\left(t;\mathbf{x}_i\right)=Y_i(t)\lambda_0(t)\exp(\beta'\mathbf{x}_i(t))$$

Donde λ_i indica el valor del riesgo para la i-ésima paciente con cáncer de mama.

X_i representa el vector de covariables en la i-ésima paciente.

Modelo Wei, Lin y Weissfeld (WLW) dado en la ecuación (2.12)

$$\boldsymbol{\lambda}_{ik}\left(t;X_{ik}\right)=Y_{ik}\left(t\right)\lambda_{0k}(t)\exp(\beta_k'\mathbf{x}_i(t))$$

Donde λ_{ik} indica el valor del riesgo para la i-ésima paciente y k-ésima recurrencia de cáncer de mama.

Donde $\lambda_{0k}(t)$ es una función de riesgo base no especificada para la k-ésima recurrencia de cáncer y $\boldsymbol{\beta}_k=(\beta_{1k},...,\beta_{pk})$ son los parámetros de regresión de evento específico o recurrencia de cáncer.

X_i representa el vector de covariables en la i-ésima paciente.

Modelo Prentice, Williams y Peterson dado en la ecuación (2.14)

$$\lambda_{ik}\left(t;X_{ik}\right)=\lambda_{0k}(t)\exp(\beta_k'X_i(t))$$

Los componentes se describen de manera similar al modelo 2.12

Modelo de Fragilidad Compartida dado en la ecuación (2.25)

$$\lambda_{ij}(t_{ij})=\lambda_o(t_{ij})\omega_i\exp(\beta'X_{ij})$$

Donde λ_{ij} indica el valor de riesgo para la i-ésima paciente y j-ésima recurrencia de cáncer de mama.

$t_{ij,}$ es el tiempo de recurrencia de cáncer de mama en la i-ésima paciente y j-ésima recurrencia.

Donde λ_0 es la función de riesgo base, ω_i es el parámetro de efecto aleatorio común que toma en cuenta la dependencia entre recurrencias sucesivos de cáncer dentro de una misma paciente.

$X_{ij} = (X_{1ij}, ..., X_{pij})$, es el vector de covariables fijas para la paciente i.

Para encontrar el modelo que mejor ajuste a los datos se utilizó como indicador de selección el Criterio de Información Akaike (AIC) dado por:

$$AIC = -2\log(l) + 2p \tag{2.32}$$

Donde p denota el número de covariables en el modelo.

2.1. Estimación de los parámetros de regresión usando el método de máxima verosimilitud parcial con el algoritmo de Newton Raphson para los modelos de eventos recurrentes sin considerar efecto aleatorio (sin fragilidad), usando la ecuación (2.15).

2.2. Estimación de los parámetros de regresión usando el método de máxima verosimilitud penalizada con el algoritmo de Marquardt para el modelo de Fragilidad Compartida de eventos recurrentes dado en la ecuación (2.30).

3. Análisis de Residuales de los modelos de eventos recurrentes sin efecto aleatorio dado en la ecuación (2.23, 2.24).
4. Interpretación y discusión de resultados.

IV. RESULTADOS Y DISCUSIÓN

4.1. Análisis exploratorio y descriptivo de las variables de estudio

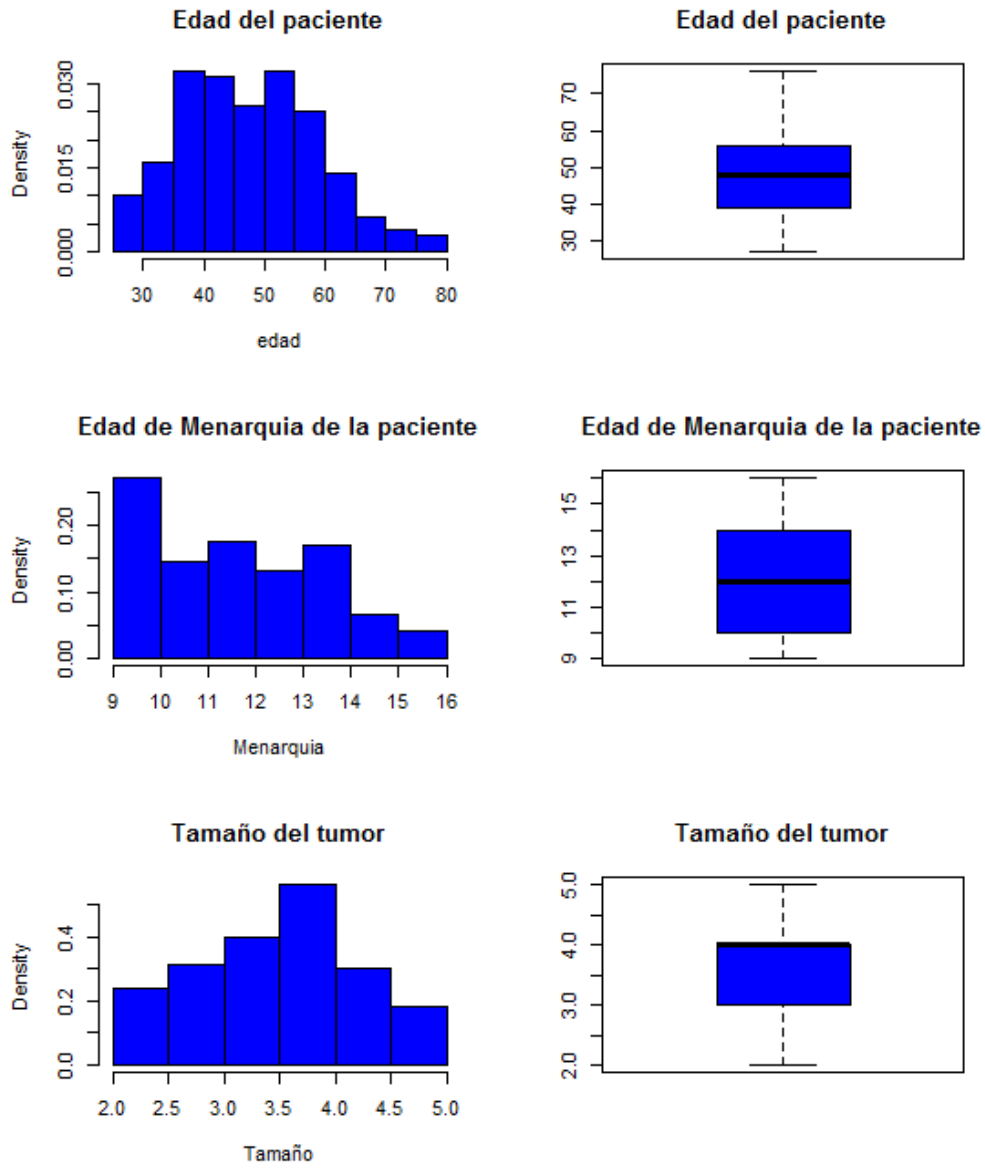


Figura 5: Distribución de las variables cuantitativas continúas de las pacientes con cáncer de mama

La Figura 5, proporciona información sobre la forma de la distribución de las variables continuas, del cual se puede observar que la edad de la paciente se centra alrededor de 35 a 55 años, y presenta un promedio de 49.49 años con una variación de 12.045 (ver Cuadro 3), mientras que la edad promedio de menarquia de la paciente es de 12.1 años y una variación de 2.029 años.

Por otro la lado, la forma de la distribución del tamaño del tumor al diagnóstico de cáncer de mama en la muestra de pacientes tiene apariencia simétrica, con promedio de 3.69 y variación de 0.7815 mm. De acuerdo al diagrama de cajas, en ninguna de las variables se observa valores extremos.

Cabe resaltar que el número máximo de recurrencia de cáncer observado en los pacientes fue de 5 y mínimo de 1, con un promedio de 1.926 recurrencias por sujeto, además el 50% de los pacientes tiene un máximo de 2 eventos de cáncer de mama. Asimismo, el tiempo promedio de recurrencia de cáncer por paciente es de 651.82 días. Debido a la alta variación en el tiempo para la ocurrencia del evento, fue necesario analizar el tiempo mediano de recurrencia; esto es, el 50% de los pacientes presenta un máximo de tiempo para el desarrollo de la enfermedad de 491.3 días (Ver Cuadro 3). Estas variables permitieron analizar la relación con el riesgo de desarrollar el cáncer de mama en pacientes con diagnóstico positivo que se presenta en la determinación de los modelos sin efecto aleatorio y de fragilidad de eventos recurrentes.

Cuadro 3: Medidas descriptivas de las variables cuantitativas de los pacientes con recurrencia de cáncer de mama

Variable	Casos	Media	Mínimo	Máximo	Desv.Est.	Mediana	Q1	Q3
Eventos	68	1.926	0	5	0.9513	2	1	2
Tiempo entre recurrencias	68	651.824	224	3200	507.695	491.3	371.13	740.25
Edad	68	49.485	27	76	12.045	49	40	57.25
Menarquia	68	12.100	9	16	2.0295	12	10	14
Tamaño	68	3.678	2	5	0.8336	4	3	4.125

Fuente: Elaboración propia

Cuadro 4: Casos de cáncer de mama en las pacientes según variables categóricas o características clínicas

Variable	Casos	Porcentaje (%)
Menopausia		
Si	37	54.4 %
No	31	45.6%
Paridad		
Multipara	46	67.6%
Nulipara	22	32.4%
Tipo Histológico		
Ductual	45	66.2%
Lobulillar	17	25.0%
Otros	6	8.8%
Grado		
I	12	17.6%
II	28	41.2%
III	28	41.2%
Compromiso Ganglionar		
Si	51	75.0%
No	17	25.0%

Fuente: Elaboración propia

El Cuadro 4, presenta las características clínicas de los casos de las pacientes con cáncer de mama. El 54.4% de los casos se encontraban en estado de menopausia cuando se les diagnosticó el cáncer, asimismo el 67.6% de las pacientes tienen de uno a más hijos, el 66.2% presenta un tipo cáncer de mama ductual, el 82.4%, se les diagnosticó cáncer de grado III y I, el 75% de las pacientes, el cáncer afectó los ganglios linfáticos de la axila en el diagnóstico. Por el alto porcentaje de casos estas variables podrían considerarse como factores asociado a que un paciente con diagnóstico positivo desarrolle una recurrencia de cáncer de mama.

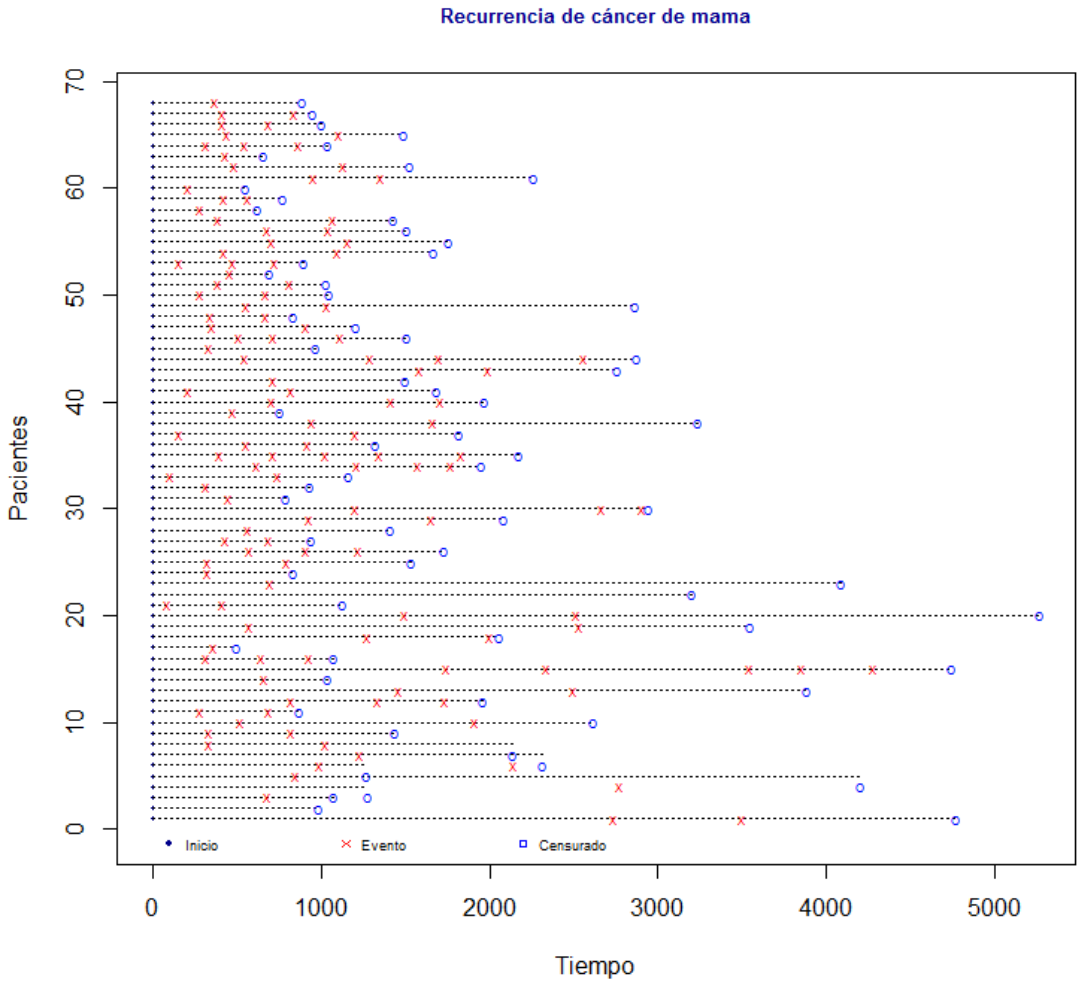


Figura 6: Gráfico del tiempo de recurrencia de cáncer de mama en pacientes del INEN. 2008-2016

La Figura 6 representa de manera exploratoria el tiempo de recurrencia de cáncer de las pacientes con diagnóstico positivo sometidas a cirugía mastectomía en el Instituto Nacional de Enfermedades Neoplásicas. Se observa que el máximo número de recurrencias de la enfermedad fue de cinco y se observó solo en una paciente. Además se registró que el tiempo máximo de seguimiento fue de 5270 días.

4.2. Determinación de los Modelos de Eventos Recurrentes

4.2.1. Modelo Andersen-Gill

Para la estimación de este modelo se usó la función de riesgo descrito en (2.7.1), cuya expresión matemática se encuentra en la ecuación (2.10). El modelo asume incrementos independientes, es decir las recurrencias de cáncer de mama son independientes dada las covariables en estudio. Como primera etapa se realizó el análisis incluyendo todas las variables registradas para la investigación, el cual se presenta en el Cuadro 5.

Cuadro 5: Resultados de la estimación del modelo Andersen-Gill

Variable	$\hat{\beta}$	$EE(\hat{\beta})$	Robusto EE	HR	IC(HR) del 95%
Edad	-0.0258***	0.0121	0.0069	0.9745	(0.9613-0.9879)
Menarquia	-0.1061**	0.0486	0.0323	0.8994	(0.8442-0.9581)
Menopausia					
Si	0.1843	0.2768	0.1379	1.2023	(0.9175-1.5755)
Paridad					
Nulipara	-0.0041	0.2082	0.1139	0.9959	(0.7966-1.2452)
Tipo Histológico					
Lobulillar	0.2377 •	0.2157	0.1307	1.2683	(0.9816-1.6388)
Otros	-0.1494	0.3685	0.1948	0.8613	(0.5879-1.2618)
Grado					
II	-0.0458	0.2519	0.1452	0.9553	(0.7187-1.2618)
III	0.0385	0.2479	0.1541	1.0392	(0.7684-1.4056)
TamañoTum	0.1196	0.1297	0.0865	1.127	(0.9512-1.3354)
Compromiso					
Si	-0.0699	0.2360	0.1274	0.9324	(0.7264-1.1969)
RL test	11.85	df=10	p=0.2956		
Wald test	11.86	df=10	p=0.0003**		
Score test	11.86	df=10	p=0.2948	Robustez=11.86	p=0.2946
AIC	1001.25				

*****<0.001, **<0.01, *<0.05, • < 0.1**

Fuente: Elaboración propia

De los resultados observados en el Cuadro 5, solo la prueba global del test de Wald (Ver ecuación 2.18) resultó significativa ($p < 0.001$), lo cual indica que al menos una de las covariables influye sobre el tiempo de recurrencia. Además, también se observa que la edad y la edad de menarquia contribuye a un nivel de significación de 1%; pero el tipo de carcinoma contribuye a un nivel del 10% sobre el tiempo de recurrencia. Posteriormente,

se realizó una selección de variables considerando el criterio de información AIC dado en la ecuación (2.32), y se obtuvo un modelo final como se muestra en el Cuadro 6.

Cuadro 6: Resultados de la estimación del modelo Andersen-Gill luego de la selección de variables

Variable	$\hat{\beta}$	$EE(\hat{\beta})$	Robusto EE	HR	IC(HR) del 95%
Edad	-0.02074***	0.08001	0.005719	0.9795	(0.9686-0.9905)
Menarquia	-0.096657**	0.046401	0.030837	0.9079	(0.8546-0.9644)
Tipo					
Lobulillar	0.202494 •	0.202718	0.114513	1.2245	(0.9783-1.5326)
Otros	-0.172681	0.33868	0.170273	0.8414	(0.6027-1.1747)
RL test	10.49	df=4	p=0.0329400		
Wald test	22.12	df=4	p=0.0001898		
Score test	10.06	df=4	p=0.0394900	Robustez=7.82	p=0.09835
AIC	990.6				

*****<0.001, **<0.01, *<0.05, • < 0.1**

Fuente: Elaboración propia

Luego de la selección de variables, el Cuadro 6 muestra que la edad afecta de manera significativa ($p < 0.001$) a la recurrencia de cáncer de mama, además la cuarta columna de la tabla presenta el riesgo estimado de 0.9795 (IC del 95%, 0.9686-0.9905). Del coeficiente estimado de la covariable edad del paciente se obtiene $1-e^{-0.02074} = 0.02052641$ el cual puede ser interpretado que por cada año que se incremente la edad, se estima que el riesgo de cáncer de mama disminuya en 2.05%, manteniendo constante la edad de su primera menstruación y el tipo histológico de cáncer. Asimismo, la edad de su primera menstruación de las pacientes es significativa ($p < 0.01$) y muestran un riesgo estimado de 0.9079 (IC del 95%, 0.8546-0.9644), y del coeficiente estimado de la covariable la edad de su primera menstruación se obtiene $1\text{-}e^{-0.096657} = 0.09213265$, esto indica que el riesgo de cáncer de mama se reduce en 9.21% por cada año, luego de desarrollar una recurrencia de cáncer del mismo tipo, manteniendo constante las demás covariables incluidas en el modelo. Por último, el riesgo estimado de recurrencia de cáncer de mama de un paciente con carcinoma lobulillar ($p < 0.1$) es $e^{0.202494} = 1.224453$ veces el riesgo de recurrencia de cáncer de mama con el carcinoma ductual, manteniendo constante las otras covariables incluídas en el modelo.

4.2.2. Forma funcional para los predictores en el modelo A-G

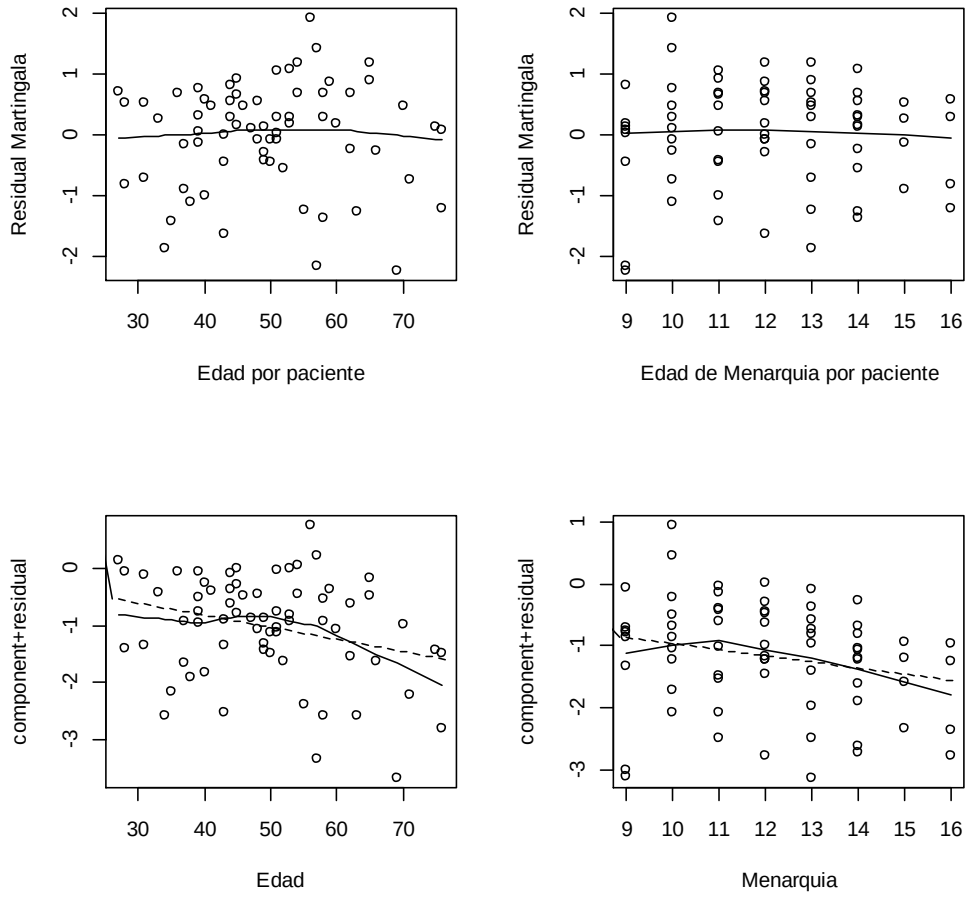


Figura 7: Forma funcional para las variables Edad, Edad de Menarquia. Modelo (A-G)

Para la evaluación de los residuales se aplicó la teoría descrita en la sección (2.6.6). Al analizar la gráfica se observa que la no linealidad parece no estar presente tanto en la variable edad del paciente y edad de su primera menarquía, por lo tanto los gráficos residuales martingala como componente más residual tienen el lowess ajustado y se aproxima a una línea recta por lo que el supuesto de linealidad del exponente de e es adecuada.

4.2.3. Diagnóstico y evaluación de la influencia de residuales para el modelo A-G

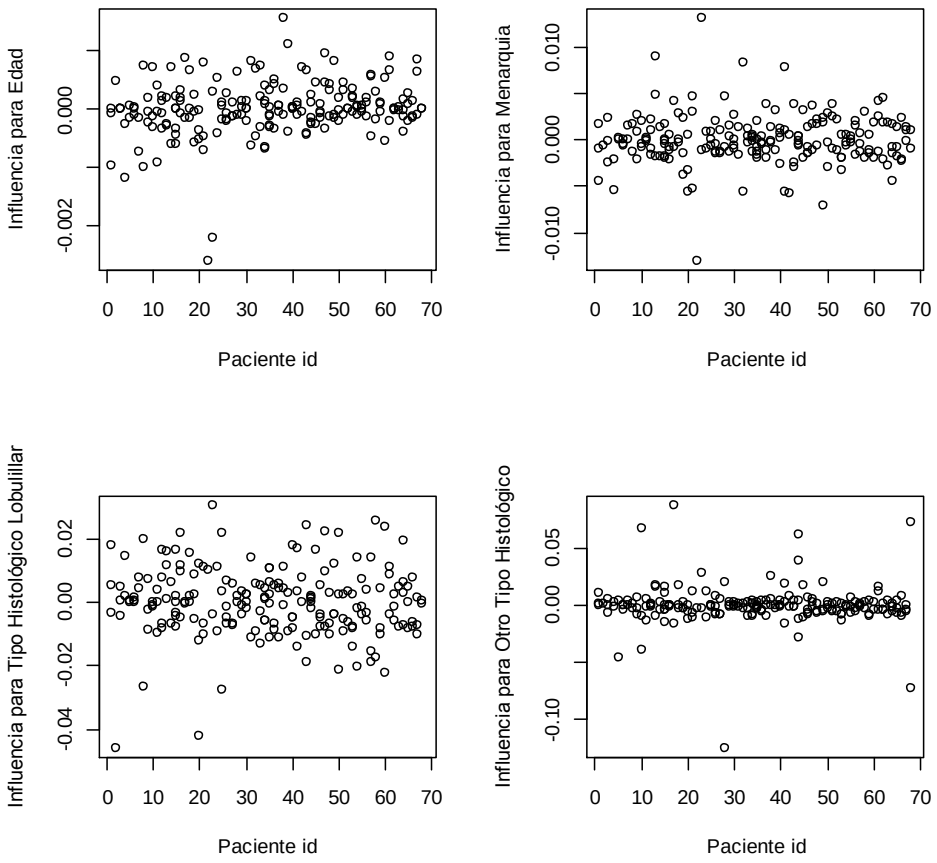


Figura 8: Influencia para las variables Edad, Menarquia, Tipo Histológico Lobulillar y otro. Modelo (A-G)

La figura 8, permitió evaluar la influencia de las observaciones sobre los coeficientes de regresión en el modelo. Para asegurar que un valor es posiblemente influyente se verificó que $\left| dfbetas_{ij} > \frac{2}{\sqrt{n}} \right|$, y como $\frac{2}{\sqrt{68}} = 0.2425$, entonces como ninguno de los dfbetas excede a 0.2425, se concluye que no hay valores influyentes en el conjunto de datos.

4.2.4. Supuesto de riesgos proporcionales para el modelo A-G

Para el supuesto de proporcionalidad se usó la teoría descrita en (3.6.7). La hipótesis planteada para la prueba de riesgos proporcionales está dado por:

Ho: Existe proporcionalidad de riesgos sobre las covariables
Hi: No existe proporcionalidad de riesgos sobre las covariables

Cuadro 7: Resultados de la asunción proporcionalidad

Variable	Chi-cuadrado	p
Edad	0.6821	0.409
Menarquia	0.0095	0.922
Tipo histológico Lobulillar	0.0496	0.824
Otro tipo histológico	0.0145	0.904
Global	0.7234	0.948

Fuente: Elaboración propia

De acuerdo a los resultados del Cuadro 7, no se ha encontrado suficiente evidencia estadística para rechazar la asunción de proporcionalidad para las variables; Edad, edad de su primera menstruación y tipo histológico. También, se puede observar que la prueba global resultó no significativa ($p>0.1$), esto indica que no se ha encontrado evidencia estadística para rechazar de que las variables cumplan con el supuesto de riesgos proporcionales. Por lo tanto se cumple con la proporcionalidad del modelo.

4.2.5. Modelo Pretince, Williams y Peterson (PWP)

Para la estimación del modelo (PWP) descrito en (2.7.2), dada en la ecuación (2.14), se asume que una paciente no está en riesgo en la k-ésima recurrencia de cáncer de mama hasta que haya experimentado una recurrencia previa. El Cuadro 8 presente la estimación del modelo.

Cuadro 8: Resultados de la estimación del modelo Pretince, Williams y Peterson (PWP)

Variable	$\hat{\beta}$	$EE(\hat{\beta})$	Robusto EE	HR	IC(HR) del 95%
Edad	-0.0426***	0.0131	0.0113	0.9583	(0.9372-0.9798)
Menarquia	-0.1833**	0.053	0.0569	0.8325|	(0.7446-0.9308)
Menopausia					
Si	0.4520	0.3059	0.2870	1.5715	(0.8954-2.7580)
Paridad					
Nulipara	-0.0681	0.2179	0.1937	0.9341	(0.6391-1.3655)
Tipo Histológico					
Lobulillar	0.6185*	0.2383	0.2809	1.8560	(1.0704-3.2185)
Otros	-0.2520	0.4043	0.3625	0.7773	(0.3819-1.5817)
Grado					
II	0.1907	0.2678	0.2769	1.2101	(0.7032-2.0823)
III	0.4395	0.2711	0.2993	1.5519	(0.8632-2.7901)
TamañoTum	0.2240	0.1419	0.1582	1.2511	(0.9176-1.7057)
Compromiso					
Si	-0.1903	0.2536	0.2469	0.8267	(0.5095-1.3415)
RL test	26.36	df=10	p=0.003		
Wald test	27.12	df=10	p=0.002		
Score test	26.59	df=10	p=0.003	Robustez=16	p=0.0843
AIC	760.28				

*****<0.001, **<0.01, *<0.05, • < 0.1**

Fuente: Elaboración propia

De los resultados observados en el Cuadro 8, las tres pruebas globales; test de Wald, Ratio de Verosimilitud (RL) y test de Score, resultaron significativas ($p < 0.001$), lo cual indica que al menos una de las covariables influye sobre el tiempo de recurrencia. Además, al igual que el modelo A-D, la edad y la edad de menarquia contribuye a un nivel de significación de 1%, y tipo de carcinoma contribuye al 5% sobre el tiempo de recurrencia. Posteriormente, se realizó la selección de variables usando la ecuación (2.32) y se obtuvo un modelo final, cuyos resultados muestran en el Cuadro 9.

Cuadro 9: Resultados de la estimación del modelo Pretince, Williams y Peterson (PWP) luego de la selección de variables

Variable	$\hat{\beta}$	$EE(\hat{\beta})$	Robusto	HR	IC(HR) del
Edad	-0.0286***	0.008623	0.00833	0.971814	(0.9561-0.9878)
Menarquia	-0.1457**	0.048593	0.04908	0.864384	(0.7851-0.9517)
Tipo Histológico					
Lobulillar	0.4484*	0.208967	0.21551	1.565822	(1.0264-2.3888)
Otros	-0.276686	0.363484	0.29622	0.758293	(0.4243-1.3551)
RL test	19.86	df=4	p=0.000		
Wald test	18.55	df=4	p=0.000		
Score test	19.1	df=4	p=0.000	Robustez=11.13	p=0.02515
AIC	754.78				

*****<0.001, **<0.01, *<0.05, • < 0.1**

Fuente: Elaboración propia

Luego de la selección de variables usando la ecuación (2.32), el Cuadro 9 muestra que la edad del paciente afecta de manera significativa ($p < 0.001$) a la recurrencia de cáncer de mama, además la cuarta columna de la tabla presenta el riesgo estimado de 0.9718 (IC del 95%, 0.9561-0.9878). Del coeficiente estimado de la covariable edad se obtiene $1-e^{-0.0286} = 0.028186$ el cual puede ser interpretado que por cada año que se incremente la edad, se estima que el riesgo de recurrir en cáncer de mama disminuye en 2.82%, manteniendo constante la edad de su primera menstruación y el tipo carcinoma. Asimismo, la edad de menarquia o su de su primera menstruación también presentó significancia estadística ($p < 0.01$), con un riesgo estimado de 0.8644 (IC del 95%, 0.7851-0.9517), del coeficiente estimado de la covariable edad de menarquia se obtiene $1-e^{-0.1457} = 0.135616$, lo cual indica que por cada año que incremente en la edad de la primera menstruación, el riesgo de desarrollar una recurrencia de cáncer disminuye en 13.56%, manteniendo constante las demás covariables. Finalmente, se estima que el riesgo de cáncer de mama de un paciente con tipo carcinoma ($p < 0.05$) es $e^{0.4484} = 1.565822$ veces el riesgo de recurrencia de cáncer de mama con el tipo carcinoma ductual, manteniendo constante las otras covariables incluídas en el modelo

4.2.6. Forma funcional para los predictores en el modelo (PWP)

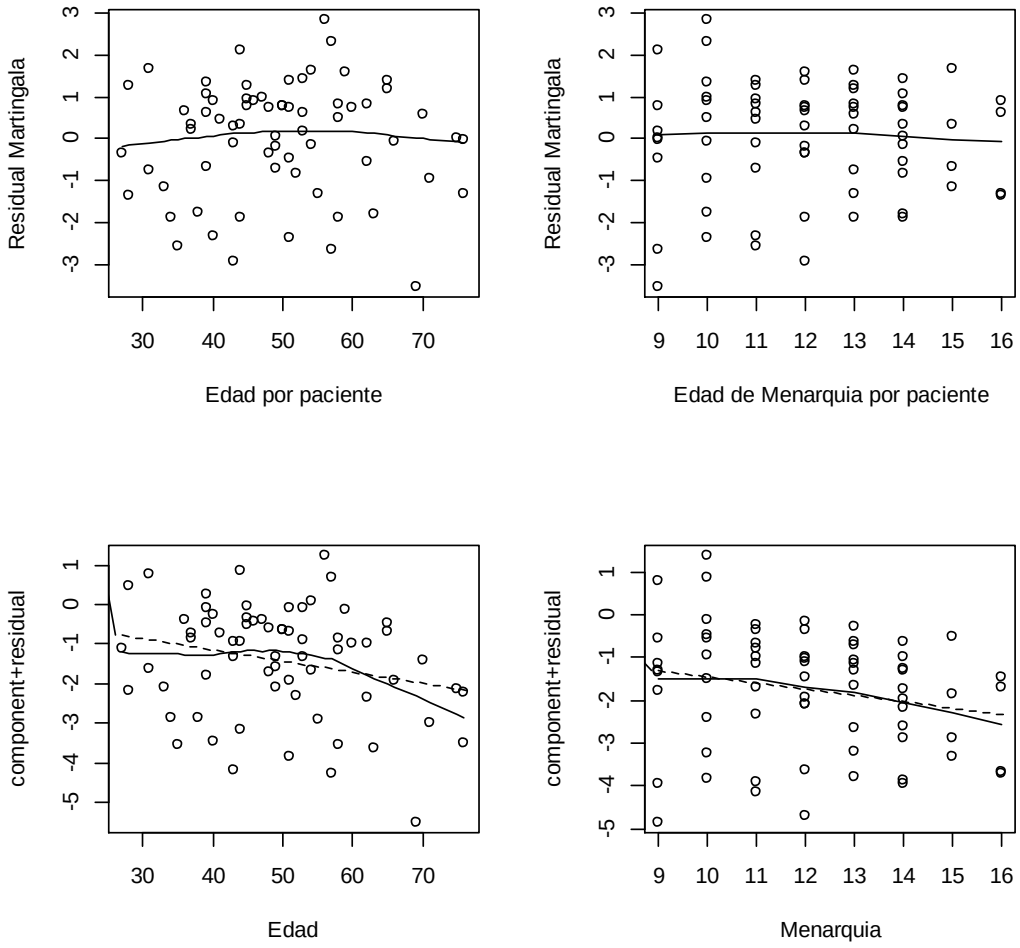


Figura 9: Forma funcional para las variables Edad, Menarquia. Modelo (PWP)

Para la evaluación de los residuales se aplicó la teoría descrita en la sección (2.6.6). Al analizar la gráfica se observa que la no linealidad parece ser ligeramente leve en la variable edad del paciente, y lo contrario sucede con la variable edad de menarquía. Por lo tanto los gráficos residuales martingala como componente más residual tienen el lowess ajustado y se aproxima a una línea recta por lo que el supuesto de linealidad del exponente e es adecuada.

4.2.7. Diagnóstico y evaluación de la influencia de residuales para el Modelo Williams y Peterson (PWP)

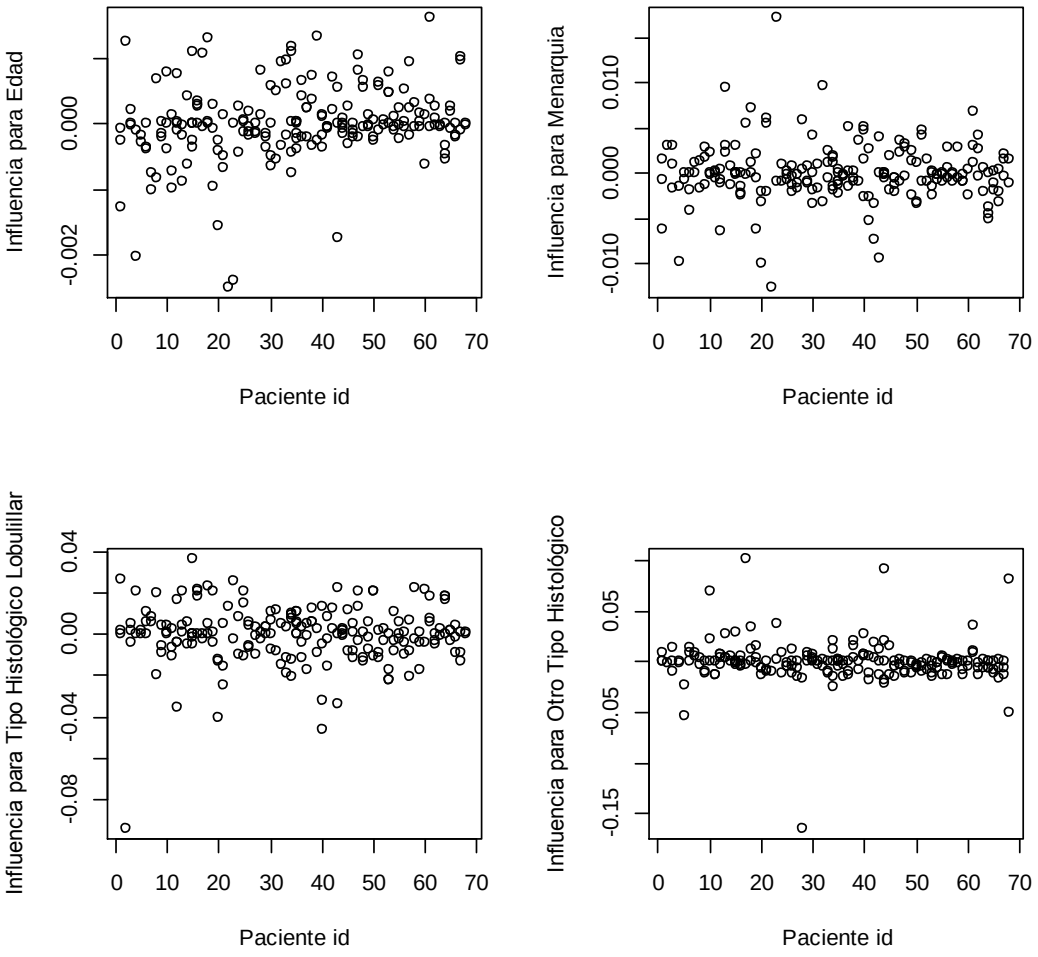


Figura 10: Influencia para las variables Edad, Menopausia, Tipo Histológico Lobulillar y Otro. Modelo (PWP)

La Figura 10 presenta la influencia de los residuales en los coeficientes de regresión, y para asegurar que un valor es posiblemente influyente se verifica que $\left| dfbetas_{ij} > \frac{2}{\sqrt{n}} \right|$

$\frac{2}{\sqrt{68}} = 0.2425$, y como ninguno de los dfbetas excede a 0.2425 entonces no hay valores influyentes en los datos.

4.2.8. Supuesto de riesgos proporcionales para el modelo (PWP)

Para el supuesto de proporcionalidad se usó la teoría descrita en (3.6.7). La hipótesis planteada para la prueba de riesgos proporcionales está dado por:

Ho: Existe proporcionalidad de riesgos sobre las covariables
Hi: No existe proporcionalidad de riesgos sobre las covariables

Cuadro 10: Resultados de la asunción proporcionalidad

Variable	Chi-cuadrado	p
Edad	3.359	0.0668
Menarquia	0.215	0.6426
Tipo histológico Lobulillar	0.433	0.5104
Otro tipo histológico	0.239	0.6249
Global	3.656	0.4546

Fuente: Elaboración propia

Según los resultados del Cuadro 10, no se ha encontrado suficiente evidencia estadística para rechazar la asunción de proporcionalidad para las variables ($p > 0.1$) para las variables edad de la primera menstruación y Tipo de carcinoma de cáncer de mama. Mientras que para la variable edad del paciente la prueba para valores mayores a 0.0668. También se puede observar que la prueba global resultó no significativa ($p>0.1$) esto indica que no se ha encontrado evidencia estadística para rechazar que las variables cumplan con el supuesto de riego proporcionales. Por lo tanto, se puede aceptar que existen riesgos proporcionales en el modelo.

4.2.9. Modelo Wei, Lin y Weissfeld (WLW)

Para la estimación del modelo (WLW) descrito en (2.7.2), dad en la ecuación (2.12), el cual asume que las recurrencias de cáncer de mama son independientes y por lo tanto no está condicionado a las recurrencias previas. Los resultados obtenidos se encuentran en el Cuadro 11.

Cuadro 11: Resultados de la estimación del modelo Wei, Lin y Weissfeld (WLW)

Variable	$\hat{\beta}$	$EE(\hat{\beta})$	Robusto EE	HR	IC(HR) del 95%
Edad	-0.0532***	0.0133	0.0131	0.9482	(0.9242-0.9728)
Menarquia	-0.2432***	0.0534	0.0679	0.7841	(0.6865-0.8957)
Menopausia					
Si	0.5712	0.3158	0.3551	1.7704	(0.8828-3.5507)
Paridad					
Nulipara	0.0680	0.2198	0.2332	1.0704	(0.6777-1.6906)
Tipo Histológico					
Lobulillar	0.8567**	0.2373	0.3146	2.3554	(1.2715-4.3633)
Otros	-0.2058	0.3939	0.4101	0.8140	(0.3644-1.8183)
Grado			|		
II	0.3651	0.2678	0.3217	1.4406	(0.7669-2.7063)
III	0.6506 •	0.2697	0.3408	1.9166	(0.9827-3.7380)
TamañoTum	0.3307 •	0.1443	0.1928	1.3920	(0.9539-2.0311)
Compromiso					
Si	-0.4239	0.2549	0.2706	0.6545	(0.3851-1.1124)
RL test	42.92	df=10	p=0.0000		
Wald test	27.82	df=10	p=0.0019		
Score test	43.6	df=10	p=0.0000	Robustez=16.22	p=0.0935
AIC	806.8				

*****<0.001, **<0.01, *<0.05, • < 0.1**

Fuente: Elaboración propia

De los resultados observados en el Cuadro 11, las tres pruebas globales; test de Wald, Ratio de Verosimilitud (RL) y test de Score, resultaron significativas ($p < 0.001$), lo cual indica que al menos una de las covariables influye sobre el tiempo de recurrencia. Además se observó resultados similares a los modelos analizados de Andersen- Gill (Ver Cuadro 5) y Pretince, Williams y Peterson (Ver Cuadro 8). El modelo de Wei, Lin y Weissfeld mostró que la edad del paciente y la edad de su primera menstruación contribuye a un nivel de 1%, Tipo carcinoma lobulillar a un nivel del 1%, Grado II del tumor y el Tamaño del tumor contribuye al 10% sobre el tiempo de recurrencia. Estas dos últimas variables difieren a los modelos analizados anteriormente en la evidencia estadística significativa. Sin embargo, posteriormente se realizó la selección de variables usando la ecuación (2.32) y se obtuvo un modelo final que se muestra en la Cuadro 12.

Cuadro 12: Resultados de la estimación del modelo Wei, Lin y Weissfeld (WLW), luego de la selección de variables.

Variable	$\hat{\beta}$	$EE(\hat{\beta})$	Robusto EE	HR	IC(HR) del 95%
Edad	-0.0350***	0.00861	0.00989	0.965560	(0.9470 -0.9843)
Menarquia	-0.1805**	0.04925	0.06215	0.834886	(0.7389-0.9426)
Tipo Histológico					
Lobulillar	0.5984*	0.2089	0.2649	1.82179	(1.0839-3.0619)
Otros	-0.1389	0.3462	0.3186	0.87025	(0.4661-1.6250)
RL test	30.08	df=4	p=4.711e-06		
Wald test	18.73	df=4	p=0.000887		
Score test	28.66	df=4	p=9.163e-05	Robustez=11.04	p=0.02627
AIC					

*****<0.001, **<0.01, *<0.05, • < 0.1**

Fuente: Elaboración propia

Luego de la selección de variables, los resultados observados en el Cuadro 12 muestra que la variable edad del paciente afecta de manera significativa ($p < 0.001$) a la recurrencia de cáncer de mama, además la cuarta columna de la tabla presenta el riesgo estimado de 0.9656 (IC del 95%, 0.9470-0.9845). Del coeficiente estimado de la covariable edad del paciente se obtiene $1-e^{-0.0350} = 0.03444$, que significa que por cada año que se incremente la edad, se estima que el riesgo de recurrir en cáncer de mama disminuya en 3.44%., manteniendo constante la edad de su primera menstruación y el tipo histológico. Además, la edad de su primera Menstruación del paciente es significativa ($p < 0.01$) con un riesgo de 0.8349 (IC del 95%, 0.7389-0.9426), y del coeficiente estimado de la covariable la edad de su primera menstruación se obtiene $1\text{-}e^{-0.1805} = 0.165114$, que indica que por cada año de inicio de menstruación el riesgo de recurrencia de cáncer disminuye en 16.51%, manteniendo constante las otras covariables consideradas en el modelo. Finalmente, se estima que el riesgo de recurrir en cáncer de mama para un paciente con tipo de carcinoma lobulillar ($p < 0.05$) es $e^{0.5984} = 1.82179$ veces el riesgo de cáncer de mama que con el carcinoma ductual, manteniendo constante las demás covariables.

4.2.10. Forma funcional para los predictores en el modelo (WLW)

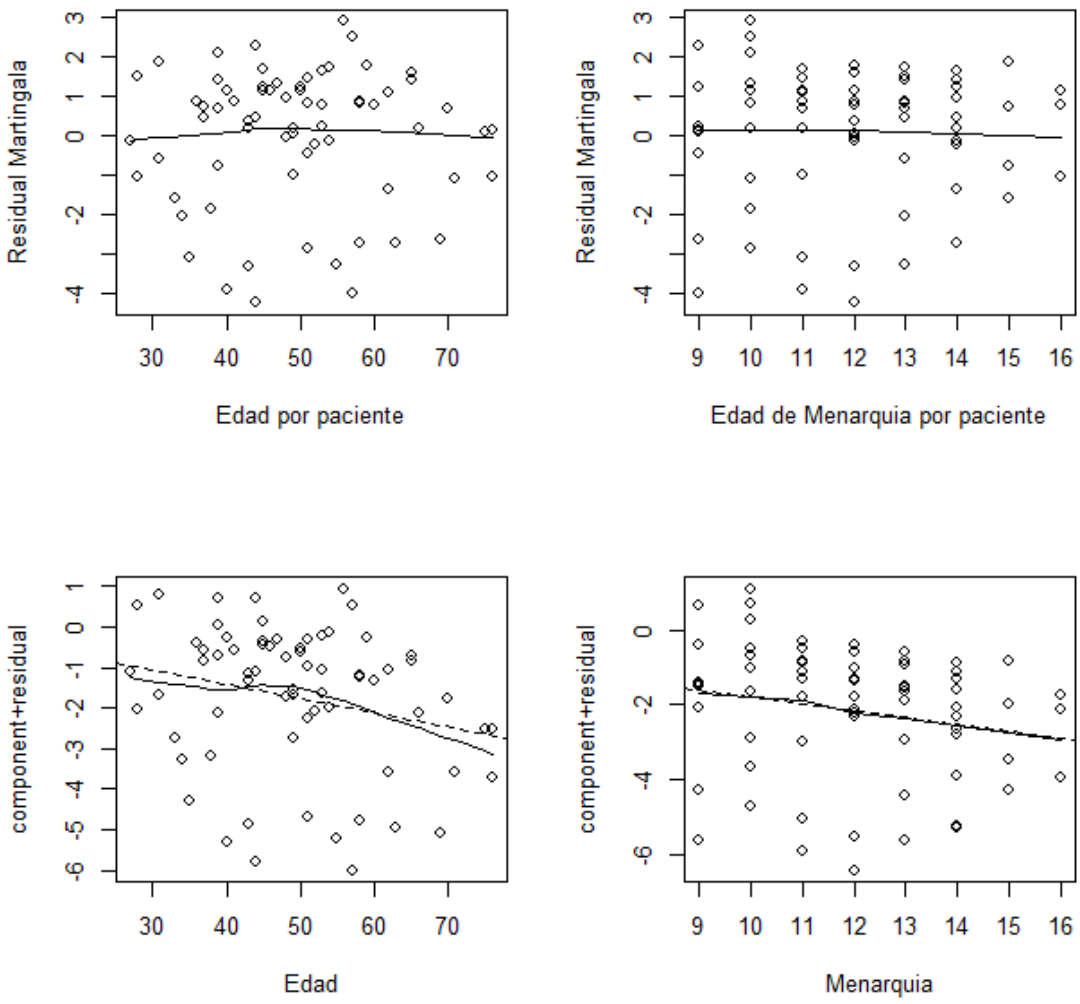


Figura 11: Forma funcional para las variables Edad, Menarquia. Modelo (WLW)

Para la evaluación de los residuales se aplicó la teoría descrita en la sección (2.6.6). Al analizar la gráfica se observa que la no linealidad parece ser ligeramente leve en la variable Edad del paciente, y en la variable Edad de su primera menstruación. Sin embargo los gráficos residuales martingala como componente más residual presentan que el lowess está ajustado y se aproxima bastante bien a una línea recta por lo que el supuesto de linealidad del exponente e es adecuada.

4.2.11. Diagnóstico y evaluación de la influencia de residuales para el Modelo Wei, Lin y Weissfeld (WLW)

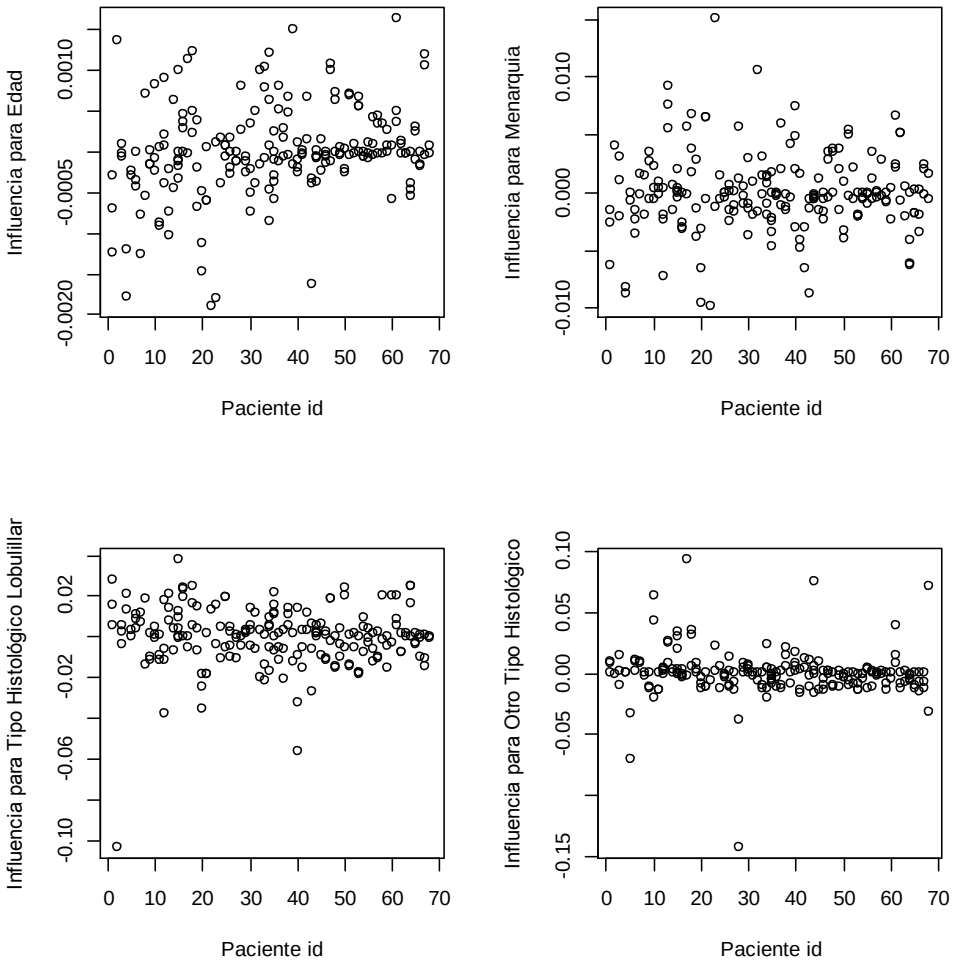


Figura 12: Influencia para las variables Edad, Menopausia, Tipo Histológico Lobulillar y Otro. Modelo (WLW)

La Figura 12 presenta la influencia de los residuales en los coeficientes de regresión, y para asegurar que un valor es posiblemente influyente se verifica que $\left| dfbetas_{ij} > \frac{2}{\sqrt{n}} \right|$, es decir $\frac{2}{\sqrt{68}} = 0.2425$ y como ninguno de los dfbetas excede a 0.2425 entonces no hay valores influyentes en los datos.

4.2.12. Supuesto de riesgos proporcionales para el modelo (WLW)

Para el supuesto de proporcionalidad se usó la teoría descrita en (3.6.7). La hipótesis planteada para la prueba de riesgos proporcionales está dado por:

Ho: Existe proporcionalidad de riesgos sobre las covariables
Hi: No existe proporcionalidad de riesgos sobre las covariables

Cuadro 13: Resultados de la asunción proporcionalidad

Variable	Chi-cuadrado	p
Edad	11.60	0.0006
Menarquia	1.08	0.2991
Tipo histológico Lobulillar	2.74	0.0976
Otro tipo histológico	1.44	0.2305
Global	14.00	0.0073

Fuente: Elaboración propia

De acuerdo a los resultados del Cuadro 13, se ha encontrado suficiente evidencia estadística para rechazar la asunción de proporcionalidad ($p < 0.001$) para la variable Edad de la paciente. Además, se puede observar que la prueba global resultó significativa, lo que indica que se ha encontrado evidencia estadística para rechazar que las covariables cumplen con el supuesto de riego proporcionales en el modelo ($p < 0.01$). Por lo tanto, para el modelo (WLW) de eventos recurrentes no se puede aceptar que existen riesgos proporcionales.

4.2.13. Modelo de Fragilidad

Se aplicó la teoría descrita en la sección 2.8, dentro del cual el modelo de fragilidad considera un efecto aleatorio no observable denominado fragilidad el cual toma en cuenta la dependencia entre las recurrencias sucesivas de cáncer de mama dentro una misma paciente. La expresión del modelo está dada en la ecuación (2.25). Los resultados obtenidos se encuentran en el Cuadro 14.

Cuadro 14: Resultados de la estimación del Modelo de Fragilidad Compartida Gamma para eventos recurrentes

Variable	$\hat{\beta}$	$EE(\hat{\beta})(H)$	$EE(\hat{\beta})$(HIH)	HR	IC(HR) del 95%
Edad	-0.0258*	0.0121	0.0121	0.9745	(0.95-1.00)
Menarquia	-0.1020*	0.0486	0.0486	0.9030	(0.82-0.99)
Menopausia					
Si	0.1907	0.2778	0.2778	1.2101	(0.68-2.09)
Paridad					
Nulipara	-0.0162	0.2083	0.2083	0.9838	(0.65-1.48)
Tipo					
Lobulillar	0.2389	0.2159	0.2159	1.2622	(0.83-1.93)
Otros	-0.1225	0.3686	0.3686	0.8847	(0.43-1.82)
Grado					
II	-0.0569	0.2514	0.2514	0.9447	(0.58-1.55)
III	0.0365	0.2475	0.2475	1.0371	(0.64-1.68)
TamañoTum	0.1198	0.1296	0.1296	1.1272	(0.87-1.45)
Compromiso					
Si	-0.0621	0.2360	0.2360	0.9398	(0.59-1.49)
Fragilidad $\hat{\theta}$	3.990e-15	SE ($\hat{\theta}$):	1.3914e-08	p=0.5	
LCV	5.1608				
Nodos usados	6				
Parámetro suavizado	3.438e+10				

*****<0.001, **<0.01, *<0.05, • < 0.1**

Fuente: Elaboración propia

De acuerdo a los resultados presentados en el Cuadro 14, los factores asociados a la recurrencia de cáncer de mama es la Edad del paciente (Riesgo relativo = 0.9745; IC al 95%, 0.95-1.00) y la Edad de su primera menstruación (Riesgo relativo = 0.9030; IC al 95%, 0.82-0.99). Sin embargo, el riesgo de desarrollar una recurrencia de cáncer condicionada a la fragilidad disminuye por cada año en la Edad en 2.55% y por cada año de Edad de inicio de la menstruación en 9.7%. Asimismo, se observó en los resultados que

el error estándar estimado del término fragilidad es S.E($\hat{\theta}$)=1.3914e-8, este valor se usa para verifica el supuesto de independencia usando la prueba de Wald unilateral en el modelo: $\hat{\theta}$/S.E($\hat{\theta}$)=3.99e-15/1.3914e-08=2.87e-07, este resultado indica que el efecto aleatorio no es significativo; por lo tanto, no se puede afirmar que exista fragilidad individual que influya en la recurrencia de cáncer de mama del paciente.

Si comparamos los resultados obtenidos con los modelos sin fragilidad podemos resaltar que tanto el modelo A-G y PWP presenta las mismas variables significativas y cumple el supuesto de riesgos proporcionales. En contraparte a los resultados encontrados con WLW que no cumple con este supuesto. El modelo de fragilidad tiene como termino de fragilidad individual no significativa; por tanto, la recurrencia de cáncer de mama en los pacientes es independiente de la fragilidad. Entonces, los modelos WLW y el modelo de Fragilidad Compartida Gamma no son adecuados para ajustar los datos.

V. CONCLUSIONES

1. Los resultados encontrados en el análisis descriptivo resalta que el 50% de los pacientes con diagnóstico positivo de cáncer de mama presenta un tiempo mediano de recurrencia de 491.3 días. A pesar de que se observó que el 75% de pacientes presenta hasta 3 recurrencia, el número máximo de recurrencia de cáncer entre los pacientes es de 5.

2. Se comparó los modelos de eventos recurrentes sin efecto aleatorio, evaluando el criterio de información AIC para la selección de variables, se encontró que con los tres modelos de Andersen-Gill (A-G), Pretince, Williams y Peterson (PWP) y Wei, Lin y Weissfeld (WLW), los factores de riesgo asociados a la recurrencia de cáncer de mama en pacientes con diagnóstico positivo están dadas por la Edad ($p < 0.001$), Edad de primera menstruación ($p<0.01$) y tipo carcinoma lobulillar ($p<0.1$)

3. Al analizar los residuales se obtuvo que solo los modelos Andersen-Gill (A-G), Pretince y Williams y Peterson (PWP) cumplen con el supuesto de riesgos proporcionales ($p >0.1$), y al comparar el nivel de precisión mediante la amplitud de cada uno de sus intervalos se encontró que el modelo A-G es más preciso por presentar menor amplitud.

4. Al analizar el modelo de fragilidad compartida gamma, se obtuvo que el término fragilidad no influye en la recurrencia de cáncer de mama; por lo tanto, prevalece el supuesto de la independencia de eventos. Además, se encontró que las variables Edad del paciente y Edad de menarquia resultaron significativas, al igual que en los dos modelos mencionados en la conclusión anterior. Sin embargo, no es un modelo adecuado para evaluar el riesgo de recurrencia de cáncer dado que la fragilidad no es un término significativo.

VI. RECOMENDACIONES

- Realizar seguimiento a grupos de pacientes con diagnóstico de cáncer haciendo uso de tratamientos especializados, seguido de un registro sistematizado de todas variables que se encuentren asociadas a la recurrencia de cáncer de mama.

- Aplicar y comparar otros modelos de eventos recurrentes como los modelos de riesgo aditivos y los modelos multiplicativos.

- Aplicar modelos de fragilidad compartido gamma de eventos recurrentes usando el algoritmo de estimación EM (Esperanza-Maximización).

- El uso y aplicación de los modelos de eventos recurrentes clásicos también pueden ser usados en otras áreas como en la Industria, Finanzas, Social entre otros, considerando la independencia de eventos.

- Aplicar y analizar modelos de fragilidad con datos de eventos recurrentes y evento de muerte.

VII. REFERENCIAS BIBLIOGRÁFICAS

Andersen, P.K; Borgan, O; Gill, R.D. 1982. Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics 10(4): 1100-1120.

Andersen, P.K; Borgan, O; Gill, R.D; and Keiding, N. 1993. Statistical Models Based on Counting Processes. 1 ed. New York. United States of America. Springer Series in Statistics, Spring-Verlag. 784p.

Androulakis, E; Koukouvinos, C; Vonta, F. 2012. Estimation and variable selection via frailty models with penalized likelihood. Statistics in Medicine 31(20): 2223–2239.

Allignol, A., Latouche, A. CRAN Task View: Survival Analysis. Consultado 16 Jul. 2016. Disponible en https://cran.r-project.org/web/views/Survival.html

American Cancer Society. About Breast Cancer (en línea, sitio web). Consultado 24 Mar. 2017. Disponible en https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html

Barbosa, R; Linás, H. 2013. Proceso Estocástico con aplicaciones. Bogotá. Colombia. Editorial Universidad del Norte. 132 p.

Barceló, MA 2002. Modelos marginales y condicionales en el análisis de supervivencia multivariante. Gaceta Sanitaria 16(2): 59-68.

Baye, F.2011. The frailty model versus the Andersen-Gill model for the prediction of recurrent events. Thesis Master of statistics: Biostatistics. Diepenbeek. Bélgica. Universiteit Hasselt. 35p

Castañeda, J y Gerritse, B. 2010. Appraisal of Several Methods to Model Time to Multiple Events per Subject: Modelling Time to Hospitalizations and Death. Revista Colombiana de Estadística 33(1): 43-61.

Centro de Prevención de Prevención de Cáncer. Cuidar de su Salud su Fuente de Información. Cáncer de Seno (en línea, sitio web). Consultado 30 Mar. 2017. Disponible en http://www.diseaseriskindex.harvard.edu/update/hccpquiz.pl?lang=spanish&func=show&quiz=breast&page=risk_list

Colosimo, R y Ruiz, S. 2006. Análise de Sobrevivéncia Aplicada. 1ed. Sao Paulo. Brasil. Editora Edgard Blucher. 369p.

Cox, D.1972. Regression models and life-tables, Journal of the Royal Statistical Society.Series B. Methodological 34(1972):187–220.

Duchateau, L. y Janssen, P .2008. Statistics for Biology and Health: The Frailty Model. ed. I. New York. United States of America. Springer Series in Statistics, Springer Science+Business Media. 328p

EI Opara, J. 2007. The interpretation and clinical application of the. word 'parity': a survey (en línea). An International Journal of Obstetrics & Gynaecology 114 (10): 1295-1297. Consultado 28 Dic. 2016. Disponible en http://onlinelibrary.wiley.com/doi/10.1111/j.1471-0528.2007.01435.x/pdf

González, JR y Peña ED, 2004. Estimación No Paramétrica de la Función de Supervivencia para Datos con Eventos Recurrentes. Rev. Esp. Salud 78(2).

Hernández, M. 2009. Epidemiología: diseño y análisis de estudios. Ed. II. Bogotá. Colombia. Editorial Médica Internacional. 406p.

Hougaard, P.1984. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. Biometrika 71(1): 75-83.

Hougaard, P.1987. Some Remarks About "Mortality of a Heterogeneous Cohort; Description and Implications". Biometrical Journal, 29 (2): 247-248.

Joly, P; Commenges, Daniel; Letenneur L. 1998. A Penalized Likelihood Approach for Arbitrarily Censored and Truncated Data: Application to Age-Specific Incidence of Dementia. Biometrics 54(1):185-194.

Kelly, PJ; Lim, LY. 2000. Survival Analysis for Recurrent Event Data: An Application To Childhood Infectious Disease. Statistics in Medicine 19: 13-33.

Kleinbaum, DG y Klein, M. 2005. Survival Analysis: A self-Learning Text. 2nd ed. Springer, Springer. New York , Inc. United States of America. 596p.

Klein, J et al. 1997. Survival analysis: Techniques for Censored and Truncated Date.

Klein, J. y Moeschberger, M., 2003. Survival analysis: Techniques for Censored and Truncated Date. 2nd ed. Springer. New York , Inc. United States of America. 542p.

Lazcano, E. *et al*. 2000. Estudios de cohorte. Metodología, sesgos y aplicación. Rev. Cielo Public Health. Salud Pública de México, Vol.42, N°3

Lee y Wenyu. 2003. Statistical Methods for Survival Data Analysis.Wiley – Interscience.Oklahoma. UnitedStates of America. 535p.

Lipsitz, S; Laird, N y Harrington, D. 1990. Using the jacknife to estimate the variance of regression estimators from repeated measures studies, Communication in Statistics. Theory and Methods, Vol. 19, N° 1, 821-845.

Marquardt, D. 1963. An algorithm for Least-Squares Estimation of Nonlinear Parameters. Journal of the Society for Industrial and Applied Mathematics 11(2): 431-441.

Martines, C. y Borges, R., 2008, Modelos de estimación no paramétrica del análisis de supervivencia con eventos recurrentes, Red de Revistas Científicas de América Latina, el Caribe, España y Portugal, 15, 86-96.

Mazroui, Y; Mathoulin-Pelissier, S; Soubeyran, P; Rondeau, V. 2012. General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data. Statistics in Medicine 31: 11-12.

Mohammed, E; Shokry, E. 2012. Menopausal symptoms and the quality of life among pre/post menopausal women from rural area in Zagazig city. Life Science Journal 9(2).

Moore, D. 2016. Applied Survival Analysis Using R. Springer International Publishing Switzerland 2016. 234p

Organización de Cáncer de Mama sin fines de lucro. Tamaño de Cáncer de mama. (en línea, sitio web). Consultado 05 Feb. 2017. Disponible en http://www.breastcancer.org/es/sintomas/diagnostico/tamano

Oxford Reference. Concice Medical Dictionary 8 ed. (en línea, sitio web). Consultado 25 Mar. 2017. Disponible en http://www.oxfordreference.com/view/10.1093/acref/9780199557141.001.0001/acref-9780199557141-e-7434?rskey=xlraE7&result=8081

Peña, E. A., Strawderman, R. L. and Hollander, M. (2001), " Nonparametric Estimation with Recurrent Event Data". Journal American Association, 96, 1299-1315.

Prentice, R. L., Williams, B.J., y Peterson, A.V. (1981). One the regression analysis of multivariate failure time data. Biometrika, 68(2): 373-379.

Ramsay J. 1988. Monotone Regression Splines in Action. Statistical Science. 3(4): 425-461.

Rondeau, V; Commenges, D; Joly, P. 2003. Maximum Penalized Likelihood Estimation in a Gamma-Frailty Model. Lifetime Data Analysis 9: 139-153.

Rondeau, V; Gonzalez, J. 2005. frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. Elsevier, Mathematical and Computer Modelling, 80 (2): 154-164.

Rondeau, V. 2010. Statistical models for recurrent events and death: Application to cancer events. Elsevier, Mathematical and Computer Modelling, 52(2010)949–95.

Rondeau, V; Mazroui, T; Gonzalez, J. 2012. frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. Statistical Software 47 (4)

Solages, M. 2013. Encyclopedia of Autism Spectrum Disorders. Springer New York Publishing. USA. 1839p. Consultado 01 Abr. 2017. Disponible en https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-1698-3_1514

Sullivan, P. M. and Cai, J. 1993, "Some Graphical Displays anda Marginal Regression analyses for Recurrent Failure Times and Time Dependent Covariates". Journal American Association, 88, 811-820.

Taylor, H y Karlin, S. 1998. An Introduction To Stochastic Modelin. 3rd. ed. California United States. 646p.

Therneau T. and Grambsch P. 2001.Modeling Survival Data: Extending the Cox Model. United States of America. 346p.

Therneau T. y Grambsch P. 2000. Modeling Survival Data: Extending the Cox Model. Minnesota. United States of America. 350p.

Vaupel, J; Manton, K; Stallard, E. 1979. The Impact of Heterogeneity in Individual Frailty on the Dinamics of Mortalilty. Demography, 16:3.

Vaupel, J; Yashin, A. 1985. Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics. The American Statistics, 39 (3): 176-185.

Wang, M. C. and Chang, S.H. 1999, Nonparametric Estimation of a Recurrent Survival Function. Journal American Association, 94, 146-153.

Wei, L. J., Lin, D.Y. and Weissfeld, L. 1989, Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions". Journal American Association, 84, 1065-1073.

Wienke, A. 2011. Frailty Models in Survival Analysis. Ed. I. New York. United States of America. Taylor and Francis Group. 298p

VIII. ANEXOS

8.1. ANEXO 1

Comandos en R para el Modelo A-G

```
> fit2<-
coxph(Surv(inicio,termino,evento)~edad+Menarquia+Menopausia+Paridad+TipoH
ist1+TipoHist2+Grado2+Grado3+TamañoTum+CompGang+cluster(id),data=cancer)

> summary(fit2)
Call:
coxph(formula = Surv(inicio, termino, evento) ~ edad + Menarquia +
    Menopausia + Paridad + TipoHist1 + TipoHist2 + Grado2 + Grado3 +
    TamañoTum + CompGang + cluster(id), data = cancer)

  n= 199, number of events= 131

               coef exp(coef)  se(coef) robust se      z Pr(>|z|)
edad       -0.025823  0.974508  0.012083  0.006949 -3.716 0.000202 ***
Menarquia  -0.106083  0.899350  0.048620  0.032277 -3.287 0.001014 **
Menopausia  0.184261  1.202329  0.276767  0.137928  1.336 0.181576
Paridad    -0.004054  0.995955  0.208242  0.113963 -0.036 0.971625
TipoHist1   0.237709  1.268341  0.215669  0.130737  1.818 0.069031 .
TipoHist2  -0.149351  0.861266  0.368473  0.194837 -0.767 0.443353
Grado2     -0.045780  0.955252  0.251915  0.145169 -0.315 0.752491
Grado3      0.038498  1.039248  0.247943  0.154070  0.250 0.802686
TamañoTum   0.119601  1.127047  0.129653  0.086534  1.382 0.166931
CompGang   -0.069988  0.932405  0.236033  0.127416 -0.549 0.582811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
           exp(coef) exp(-coef) lower .95 upper .95
edad          0.9745     1.0262    0.9613    0.9879
Menarquia     0.8994     1.1119    0.8442    0.9581
Menopausia    1.2023     0.8317    0.9175    1.5755
Paridad       0.9960     1.0041    0.7966    1.2452
TipoHist1     1.2683     0.7884    0.9816    1.6388
TipoHist2     0.8613     1.1611    0.5879    1.2618
Grado2        0.9553     1.0468    0.7187    1.2697
Grado3        1.0392     0.9622    0.7684    1.4056
TamañoTum     1.1270     0.8873    0.9512    1.3354
CompGang      0.9324     1.0725    0.7264    1.1969

Concordance= 0.575  (se = 0.028 )
Rsquare= 0.058   (max possible= 0.993 )
Likelihood ratio test= 11.85  on 10 df,   p=0.2956
Wald test            = 32.45  on 10 df,   p=0.0003374
Score (logrank) test = 11.86  on 10 df,   p=0.2948,   Robust = 11.86
p=0.2946

fit2.2<-
coxph(Surv(inicio,termino,evento)~edad+Menarquia+TipoHist1+TipoHist2+clus
ter(id),data=cancer)

summary(fit2.2)
Call:
coxph(formula = Surv(inicio, termino, evento) ~ edad + Menarquia +
    TipoHist1 + TipoHist2 + cluster(id), data = cancer)

  n= 199, number of events= 131

               coef exp(coef)  se(coef) robust se      z Pr(>|z|)
edad      -0.020744  0.979469  0.008001  0.005719 -3.627 0.000286 ***
Menarquia -0.096657  0.907867  0.046401  0.030837 -3.134 0.001722 **
TipoHist1  0.202494  1.224453  0.202718  0.114513  1.768 0.077010 .
TipoHist2 -0.172681  0.841406  0.338677  0.170273 -1.014 0.310515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
edad         0.9795     1.0210    0.9686    0.9905
Menarquia    0.9079     1.1015    0.8546    0.9644
```

```
TipoHist1    1.2245     0.8167    0.9783    1.5326
TipoHist2    0.8414     1.1885    0.6027    1.1747

Concordance= 0.564  (se = 0.028 )
Rsquare= 0.051   (max possible= 0.993 )
Likelihood ratio test= 10.49  on 4 df,   p=0.03294
Wald test            = 22.12  on 4 df,   p=0.0001898
Score (logrank) test = 10.06  on 4 df,   p=0.03949,   Robust = 7.82
p=0.09835

residual<-resid(fit2.2,type='dfbeta')

residual

plot(cancer[,1],residual[,2],xlab='Paciente id',ylab='Influencia para
Menarquia')

plot(cancer[,1],residual[,5],xlab='Paciente id',ylab='Influencia para
Tipo Histológico Lobulillar')

plot(cancer[,1],residual[,6],xlab='Paciente id',ylab='Influencia para
Otro Tipo Histológico')

ri<-resid(fit2.2,type="martingale") ##Residual martingala

ri

par(mfrow=c(3,2))

plot(cancer$edad,ri,xlab="Edad",ylab="Residual Martingala")

lines(lowess(cancer$edad,ri))

edad1<-cancer[,7]

edad1<-cancer[,7]

Id<-cancer[,1]

data<-data.frame(Id,edad1,ri)

cancer$ri<-data$ri

cancer

data<-data.frame(Id,edad1,ri)


D<-rep(0,68)

for(i in 1:68)
```

```
{Dat<-subset(cancer,cancer[,1]==i)
D[i]<-sum(Dat[,17])}
D
E<-rep(0,68)
for(i in 1:68)
{Dat<-subset(cancer,cancer[,1]==i)
E[i]<-min(Dat[,7])}
E
M<-rep(0,68)
for(i in 1:68)
{Dat<-subset(cancer,cancer[,1]==i)
M[i]<-min(Dat[,8])}
M
par(mfrow=c(2,2))
plot(E,D,xlab="Edad por paciente",ylab="Residual Martingala")
lines(lowess(cancer$edad,ri))
plot(M,D,xlab="Edad de Menarquia por paciente",ylab="Residual
Martingala")
lines(lowess(cancer$Menarquia,ri))

Selección<-data.frame(E,M)

X<-as.matrix(Selección[,c("E","M")])
b<-coef(fit2.2)[c(1,2)]
for(j in 1:2){
plot(X[,j],b[j]*X[,j]+D,
xlab=c("Edad","Menarquia")[j],
ylab="component+residual")
abline(lm(b[j]*X[,j] + D ~ X[,j]), lty=2)
lines(lowess(X[,j], b[j]*X[,j] + D, iter=0)) }
```

8.2. ANEXO 2

Comandos en R para el Modelo PWP

```
##Condicional I, proceso de conteo

> fit3<-
coxph(Surv(inicio,termino,evento)~edad+Menarquia+Menopausia+Paridad+TipoH
ist1+TipoHist2+Grado2+Grado3+TamañoTum+CompGang+cluster(id)+strata(enum),
data=cancer)

> summary(fit3)
Call:
coxph(formula = Surv(inicio, termino, evento) ~ edad + Menarquia +
    Menopausia + Paridad + TipoHist1 + TipoHist2 + Grado2 + Grado3 +
    TamañoTum + CompGang + cluster(id) + strata(enum), data = cancer)

  n= 199, number of events= 131

               coef exp(coef) se(coef) robust se      z Pr(>|z|)
edad       -0.04260   0.95829  0.01312   0.01134 -3.758 0.000172 ***
Menarquia  -0.18328   0.83254  0.05281   0.05695 -3.218 0.001289 **
Menopausia  0.45200   1.57145  0.30598   0.28700  1.575 0.115278
Paridad    -0.06812   0.93415  0.21794   0.19370 -0.352 0.725067
TipoHist1   0.61845   1.85604  0.23834   0.28085  2.202 0.027662 *
TipoHist2  -0.25199   0.77725  0.40425   0.36252 -0.695 0.486986
Grado2      0.19069   1.21009  0.26779   0.27693  0.689 0.491066
Grado3      0.43948   1.55191  0.27113   0.29929  1.468 0.141990
TamañoTum   0.22401   1.25109  0.14193   0.15816  1.416 0.156656
CompGang   -0.19027   0.82674  0.25358   0.24698 -0.770 0.441076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
edad          0.9583     1.0435    0.9372    0.9798
Menarquia     0.8325     1.2011    0.7446    0.9308
Menopausia    1.5715     0.6364    0.8954    2.7580
Paridad       0.9341     1.0705    0.6391    1.3655
TipoHist1     1.8560     0.5388    1.0704    3.2185
TipoHist2     0.7773     1.2866    0.3819    1.5817
Grado2        1.2101     0.8264    0.7032    2.0823
Grado3        1.5519     0.6444    0.8632    2.7901
TamañoTum     1.2511     0.7993    0.9176    1.7057
```

```
CompGang      0.8267     1.2096    0.5095    1.3415
Concordance= 0.606  (se = 0.047 )
Rsquare= 0.124   (max possible= 0.979 )
Likelihood ratio test= 26.36  on 10 df,   p=0.003283
Wald test            = 27.12  on 10 df,   p=0.002491
Score (logrank) test = 26.59  on 10 df,   p=0.003023,   Robust = 16.58
p=0.08425

> fit3.1<-
coxph(Surv(inicio,termino,evento)~edad+Menarquia+TipoHist1+TipoHist2+clus
ter(id)+strata(enum),data=cancer)

> summary(fit3.1)
Call:
coxph(formula = Surv(inicio, termino, evento) ~ edad + Menarquia +
    TipoHist1 + TipoHist2 + cluster(id) + strata(enum), data = cancer)

  n= 199, number of events= 131

               coef exp(coef)  se(coef) robust se      z Pr(>|z|)
edad      -0.028591  0.971814  0.008623  0.008334 -3.430 0.000602 ***
Menarquia -0.145738  0.864384  0.048593  0.049086 -2.969 0.002987 **
TipoHist1  0.448411  1.565822  0.208967  0.215511  2.081 0.037463 *
TipoHist2 -0.276686  0.758293  0.363484  0.296223 -0.934 0.350281
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
edad         0.9718     1.0290    0.9561    0.9878
Menarquia    0.8644     1.1569    0.7851    0.9517
TipoHist1    1.5658     0.6386    1.0264    2.3888
TipoHist2    0.7583     1.3188    0.4243    1.3551

Concordance= 0.586  (se = 0.047 )
Rsquare= 0.095   (max possible= 0.979 )
Likelihood ratio test= 19.86  on 4 df,   p=0.0005318
Wald test            = 18.55  on 4 df,   p=0.000965
Score (logrank) test = 19.1   on 4 df,   p=0.0007525,   Robust = 11.13
p=0.02515
```

```
##Residual de influencia

par(mfrow=c(2,2))

resi<-resid(fit3.1,type='dfbeta')

resi

plot(cancer[,1],resi[,1],xlab='Paciente id',ylab='Influencia para Edad')

plot(cancer[,1],resi[,2],xlab='Paciente     id',ylab='Influencia     para
Menarquia')

plot(cancer[,1],resi[,3],xlab='Paciente  id',ylab='Influencia  para  Tipo
Histológico Lobulillar')

plot(cancer[,1],resi[,4],xlab='Paciente  id',ylab='Influencia  para  Otro
Tipo Histológico')

##Residual de Forma funcional

ri2<-resid(fit3.1,type="martingale") ##Residual martingala

ri2

edad2<-cancer[,7]

edad2<-cancer[,7]

Id2<-cancer[,1]

data2<-data.frame(Id2,edad2,ri2)

cancer$ri2<-data2$ri2

cancer[1:5,]

D2<-rep(0,68)

for(i in 1:68)

{Dat<-subset(cancer,cancer[,1]==i)

D2[i]<-sum(Dat[,17])}

D2

E2<-rep(0,68)

for(i in 1:68)

{Dat<-subset(cancer,cancer[,1]==i)

E2[i]<-min(Dat[,7])}

E2

M2<-rep(0,68)
```

```
for(i in 1:68)

{Dat<-subset(cancer,cancer[,1]==i)

M2[i]<-min(Dat[,8])}

M2

par(mfrow=c(2,2))

plot(E2,D2,xlab="Edad por paciente",ylab="Residual Martingala")

lines(lowess(cancer$edad,ri2))

plot(M2,D2,xlab="Edad de Menarquia por paciente",ylab="Residual
Martingala")

lines(lowess(cancer$Menarquia,ri2))

Selección2<-data.frame(E2,M2)

Selección2

X<-as.matrix(Selección2[,c("E2","M2")])

b<-coef(fit3.1)[c(1,2)]

for(j in 1:2){

plot(X[,j],b[j]*X[,j]+D2,

xlab=c("Edad","Menarquia")[j],

ylab="component+residual")

abline(lm(b[j]*X[,j] + D2 ~ X[,j]), lty=2)

lines(lowess(X[,j], b[j]*X[,j] + D2, iter=0)) }


##Condicional II, gap time


fit3.3<-coxph(Surv(termino-
inicio,evento)~edad+Menarquia+Menopausia+Paridad+TipoHist1+TipoHist2+Grad
o2+Grado3+TamañoTum+CompGang+cluster(id)+strata(enum),data=cancer)
summary(fit3.3)
> fit3.31<-coxph(Surv(termino-
inicio,evento)~edad+Menarquia+TipoHist1+TipoHist2+cluster(id)+strata(enum
),data=cancer)
> summary(fit3.31)
Call:
coxph(formula = Surv(termino - inicio, evento) ~ edad + Menarquia +
    TipoHist1 + TipoHist2 + cluster(id) + strata(enum), data = cancer)
```

```
  n= 199, number of events= 131

               coef exp(coef)  se(coef) robust se      z Pr(>|z|)
edad      -0.031484  0.969007  0.008517  0.008891 -3.541 0.000398 ***
Menarquia -0.164290  0.848496  0.048855  0.053009 -3.099 0.001940 **
TipoHist1  0.419062  1.520535  0.210087  0.210426  1.991 0.046427 *
TipoHist2 -0.423771  0.654574  0.360080  0.258205 -1.641 0.100752
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
edad         0.9690     1.0320    0.9523    0.9860
Menarquia    0.8485     1.1786    0.7648    0.9414
TipoHist1    1.5205     0.6577    1.0067    2.2967
TipoHist2    0.6546     1.5277    0.3946    1.0858

Concordance= 0.611  (se = 0.049 )
Rsquare= 0.112   (max possible= 0.985 )
Likelihood ratio test= 23.56  on 4 df,   p=9.772e-05
Wald test            = 21.37  on 4 df,   p=0.0002671
Score (logrank) test = 22.48  on 4 df,   p=0.000161,   Robust = 12.02
p=0.0172

  (Note: the likelihood ratio and score tests assume independence of
     observations within a cluster, the Wald and robust score tests do
not).
```

8.3. ANEXO 3

Comandos en R para el Modelo WLW

```
> fit4<-
coxph(Surv(termino,evento)~edad+Menarquia+Menopausia+Paridad+TipoHist1+Ti
poHist2+Grado2+Grado3+TamañoTum+CompGang+cluster(id)+strata(enum),data=ca
ncer)
> summary(fit4)
Call:
coxph(formula = Surv(termino, evento) ~ edad + Menarquia + Menopausia +
    Paridad + TipoHist1 + TipoHist2 + Grado2 + Grado3 + TamañoTum +
    CompGang + cluster(id) + strata(enum), data = cancer)

  n= 199, number of events= 131

               coef exp(coef) se(coef) robust se      z Pr(>|z|)
edad       -0.05330   0.94810  0.01330   0.01310 -4.069 4.72e-05 ***
Menarquia  -0.24369   0.78373  0.05342   0.06780 -3.594 0.000326 ***
Menopausia  0.56853   1.76566  0.31605   0.35530  1.600 0.109567
Paridad     0.06379   1.06587  0.21988   0.23348  0.273 0.784698
TipoHist1   0.85978   2.36264  0.23735   0.31500  2.729 0.006343 **
TipoHist2  -0.18051   0.83485  0.39050   0.40929 -0.441 0.659195
Grado2      0.36022   1.43364  0.26780   0.32152  1.120 0.262563
Grado3      0.64393   1.90395  0.26938   0.34133  1.887 0.059226 .
TamañoTum   0.33267   1.39469  0.14434   0.19273  1.726 0.084328 .
CompGang   -0.42529   0.65358  0.25515   0.27096 -1.570 0.116517
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
edad          0.9481     1.0547    0.9241    0.9728
Menarquia     0.7837     1.2759    0.6862    0.8951
Menopausia    1.7657     0.5664    0.8800    3.5427
Paridad       1.0659     0.9382    0.6745    1.6844
TipoHist1     2.3626     0.4233    1.2743    4.3805
TipoHist2     0.8348     1.1978    0.3743    1.8621
Grado2        1.4336     0.6975    0.7634    2.6923
Grado3        1.9040     0.5252    0.9752    3.7171
TamañoTum     1.3947     0.7170    0.9559    2.0348
CompGang      0.6536     1.5300    0.3843    1.1116
```

```
Concordance= 0.629  (se = 0.05 )
Rsquare= 0.195   (max possible= 0.985 )
Likelihood ratio test= 43.17  on 10 df,   p=4.633e-06
Wald test            = 27.75  on 10 df,   p=0.001977
Score (logrank) test = 43.79  on 10 df,   p=3.584e-06,   Robust = 16.2
p=0.0941

> fit4.1<-
coxph(Surv(termino,evento)~edad+Menarquia+TipoHist1+TipoHist2+cluster(id)
+strata(enum),data=cancer)
> summary(fit4.1)

Call:
coxph(formula = Surv(termino, evento) ~ edad + Menarquia + TipoHist1 +
    TipoHist2 + cluster(id) + strata(enum), data = cancer)

  n= 199, number of events= 131

               coef exp(coef)  se(coef) robust se      z Pr(>|z|)
edad      -0.035189  0.965423  0.008594  0.009886 -3.560 0.000371 ***
Menarquia -0.180842  0.834567  0.049222  0.062115 -2.911 0.003598 **
TipoHist1  0.599820  1.821790  0.208939  0.264917  2.264 0.023563 *
TipoHist2 -0.138978  0.870247  0.346215  0.318615 -0.436 0.662696
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
edad         0.9654     1.0358    0.9469    0.9843
Menarquia    0.8346     1.1982    0.7389    0.9426
TipoHist1    1.8218     0.5489    1.0839    3.0619
TipoHist2    0.8702     1.1491    0.4661    1.6250

Concordance= 0.611  (se = 0.05 )
Rsquare= 0.14   (max possible= 0.985 )
Likelihood ratio test= 30.08  on 4 df,   p=4.711e-06
Wald test            = 18.73  on 4 df,   p=0.0008869
Score (logrank) test = 28.66  on 4 df,   p=9.162e-06,   Robust = 11.03
p=0.02627
par(mfrow=c(2,2))
r<-resid(fit4.1,type='dfbeta')
```

```
r
plot(cancer[,1],r[,1],xlab='Paciente id',ylab='Influencia para Edad')
plot(cancer[,1],r[,2],xlab='Paciente id',ylab='Influencia para
Menarquia')
plot(cancer[,1],r[,3],xlab='Paciente id',ylab='Influencia para Tipo
Histológico Lobulillar')
plot(cancer[,1],r[,4],xlab='Paciente id',ylab='Influencia para Otro Tipo
Histológico')

ri3<-resid(fit4.1,type="martingale") ##Residual martingala
ri3
edad3<-cancer[,7]
edad3<-cancer[,7]
Id3<-cancer[,1]
data3<-data.frame(Id3,edad3,ri3)
cancer$ri3<-data3$ri3
cancer[1:5,]

D3<-rep(0,68)
for(i in 1:68)
{Dat<-subset(cancer,cancer[,1]==i)
D3[i]<-sum(Dat[,17])}
D3

E3<-rep(0,68)
for(i in 1:68)
{Dat<-subset(cancer,cancer[,1]==i)
E3[i]<-min(Dat[,7])}
E3

M3<-rep(0,68)
for(i in 1:68)
{Dat<-subset(cancer,cancer[,1]==i)
M3[i]<-min(Dat[,8])}
M3

par(mfrow=c(2,2))
plot(E3,D3,xlab="Edad por paciente",ylab="Residual Martingala")
lines(lowess(cancer$edad,ri3))
plot(M3,D3,xlab="Edad de Menarquia por paciente",ylab="Residual
Martingala")
```

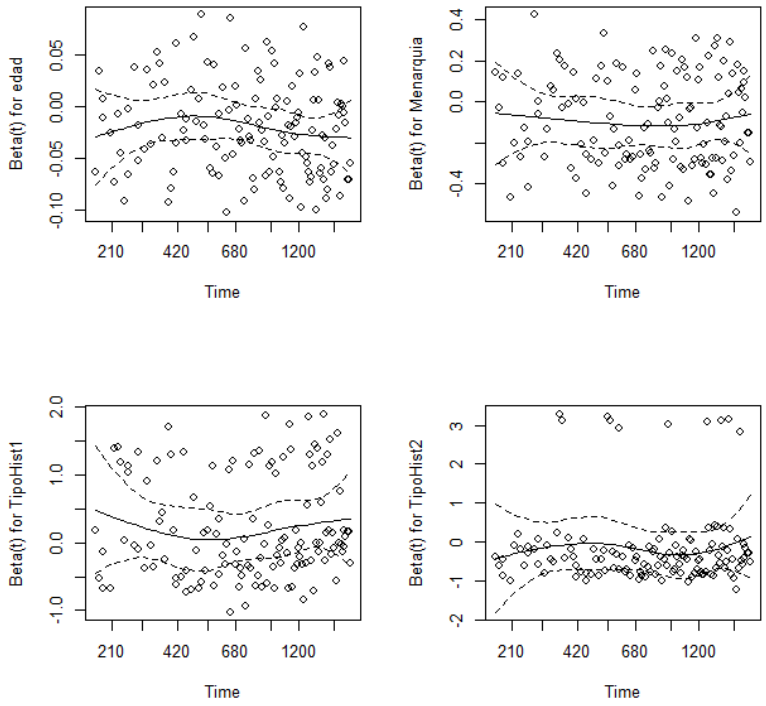
```
lines(lowess(cancer$Menarquia,ri3))

Selección3<-data.frame(E3,M3)
Selección3
X<-as.matrix(Selección3[,c("E3","M3")])
b<-coef(fit4.1)[c(1,2)]
for(j in 1:2){
plot(X[,j],b[j]*X[,j]+D3,
xlab=c("Edad","Menarquia")[j],
ylab="component+residual")
abline(lm(b[j]*X[,j] + D3 ~ X[,j]), lty=2)
lines(lowess(X[,j], b[j]*X[,j] + D3, iter=0)) }
```

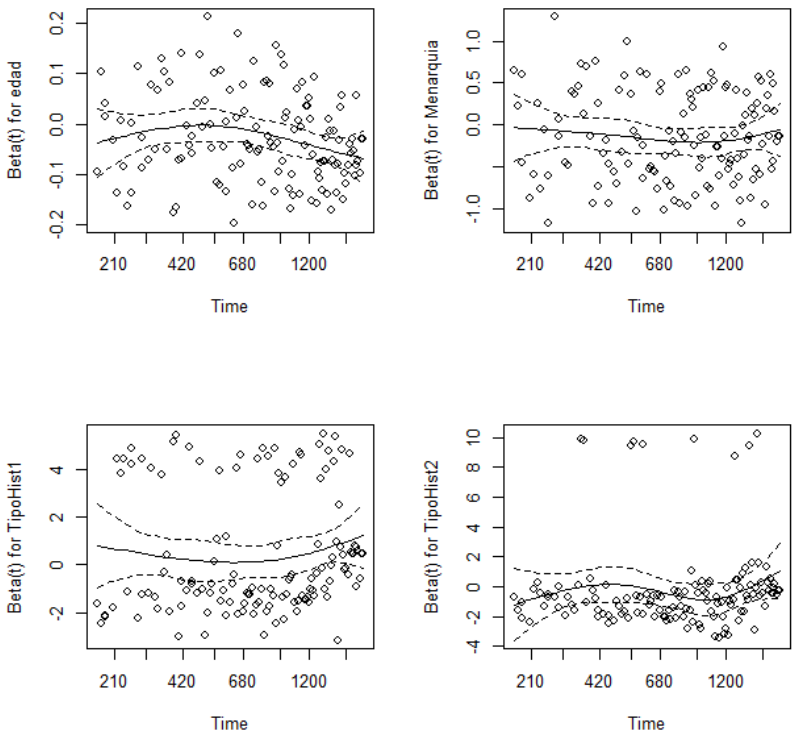
8.4. ANEXO 4

Gráfica para el supuesto de riesgos proporcionales

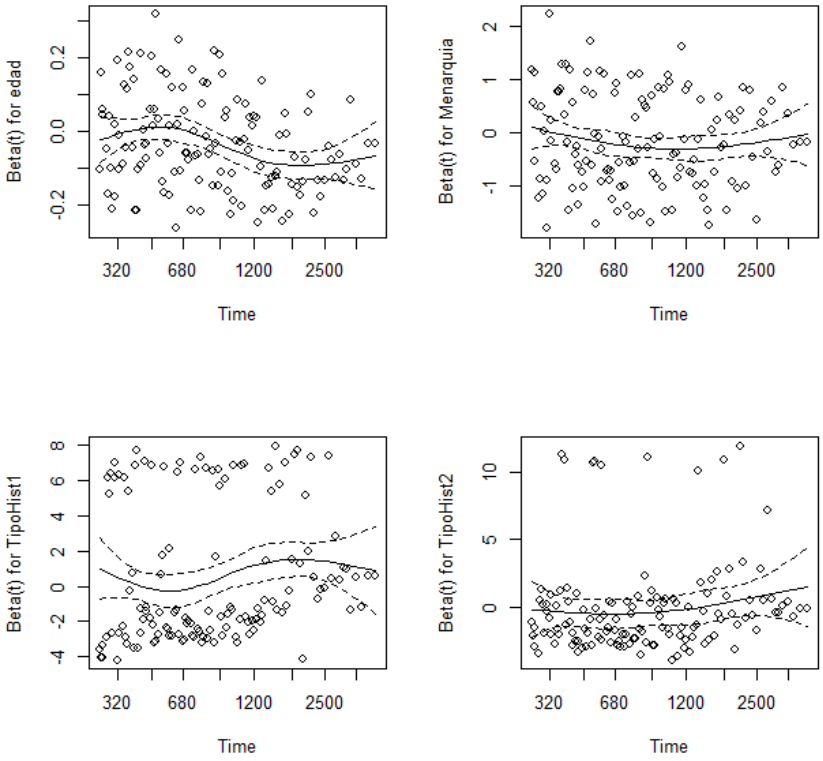
Modelo A-D



Modelo PWP



Modelo WLW


Beta(t) for edad
-0.2
0.0
0.2
320
680
1200
2500
Time
Beta(t) for Menarquia
-1
0
1
2
320
680
1200
2500
Time
Beta(t) for TipoHist1
-4
-2
0
2
4
6
8
320
680
1200
2500
Time
Beta(t) for TipoHist2
0
5
10
320
680
1200
2500
Time

8.5. ANEXO 5

Modelo de Fragilidad Gamma Compartida

```
> modfrailAG<-
frailtyPenal(Surv(inicio,termino,evento)~edad+Menarquia+as.factor(Menopau
sia)+as.factor(Paridad)+as.factor(TipoHist1)+as.factor(TipoHist2)+as.fact
or(Grado2)+as.factor(Grado3)+TamañoTum+CompGang+cluster(id),data=cancer,n
.knots=6,kappa=1000,cross.validation=TRUE,recurrentAG=TRUE)

Be patient. The program is computing ...

The program took 0.33 seconds

> modfrailAG
Call:
frailtyPenal(formula = Surv(inicio, termino, evento) ~ edad +
    Menarquia + as.factor(Menopausia) + as.factor(Paridad) +
    as.factor(TipoHist1) + as.factor(TipoHist2) +
as.factor(Grado2) +
    as.factor(Grado3) + TamañoTum + CompGang + cluster(id), data =
cancer,
    recurrentAG = TRUE, cross.validation = TRUE, n.knots = 6,
    kappa = 1000)


      Calendar timescale


  Shared Gamma Frailty model parameter estimates
  using a Penalized Likelihood on the hazard function

      coef exp(coef) SE coef (H) SE coef (HIH)        z        p
edad        -0.0258367  0.974494   0.0121496     0.0121496 -2.1265543 0.033457
Menarquia   -0.1020275  0.903005   0.0485834     0.0485834 -2.1000500 0.035724
Menopausia1  0.1907514  1.210159   0.2778132     0.2778132  0.6866174 0.492320
Paridad1    -0.0162386  0.983893   0.2082879     0.2082879 -0.0779622 0.937860
TipoHist11   0.2328830  1.262234   0.2158603     0.2158603  1.0788601 0.280650
TipoHist21  -0.1224800  0.884724   0.3685813     0.3685813 -0.3323013 0.739660
Grado21     -0.0568735  0.944714   0.2513909     0.2513909 -0.2262352 0.821020
Grado31      0.0364533  1.037126   0.2474915     0.2474915  0.1472912 0.882900
TamañoTum    0.1197764  1.127245   0.1295609     0.1295609  0.9244798 0.355240
CompGang    -0.0620803  0.939807   0.2360159     0.2360159 -0.2630346 0.792520
```

```
    Frailty parameter, Theta: 3.99043e-15 (SE (H): 1.39141e-08 ) p
= 0.5

      penalized marginal log-likelihood = -1008.01
      Convergence criteria:
      parameters = 1.54e-05 likelihood = 1.23e-06 gradient =
8.71e-11

      LCV = the approximate likelihood cross-validation criterion
            in the semi parametrical case     = 5.16081

      n= 199
      n events= 131  n groups= 68
      number of iterations:  23

      Exact number of knots used:  6
      Best smoothing parameter estimated by
      an approximated Cross validation:  3438313828, DoF:  7.00

> summary(modfrailAG)
                hr     95%     C.I.
       edad   0.97 (   0.95 -   1.00 )
   Menarquia   0.90 (   0.82 -   0.99 )
 Menopausia1   1.21 (   0.70 -   2.09 )
    Paridad1   0.98 (   0.65 -   1.48 )
  TipoHist11   1.26 (   0.83 -   1.93 )
  TipoHist21   0.88 (   0.43 -   1.82 )
     Grado21   0.94 (   0.58 -   1.55 )
     Grado31   1.04 (   0.64 -   1.68 )
   TamañoTum   1.13 (   0.87 -   1.45 )
    CompGang   0.94 (   0.59 -   1.49 )
```

8.6. ANEXO 6

Gráfico de función de riesgo para diferentes tamaño de nudos