UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACÍON EN ESTADÍSTICA APLICADA

ALGUNOS MÉTODOS ESTADÍSTICOS PARA EL ESTUDIO DE EVENTOS RECURRENTES

TESIS
QUE PARA OPTAR POR EL GRADO DE
MAESTRA EN MATEMÁTICAS

PRESENTA:
Brenda Lambert Lamazares

DIRECTORA:
Dra. Silvia Ruiz-Velasco Acosta
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)

Ciudad de México, Octubre, 2021

A mi mamá,
por hacer que 1786.48 kilómetros
se sientan como pocos metros.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología por el apoyo económico brindado durante mis estudios de posgrado.

A la Universidad Nacional Autónoma de México y en especial al Posgrado en Ciencias Matemáticas por haber contribuido a mi formación como matemática y como persona.

A la Dra. Silvia Ruiz Velasco Acosta por haberme apoyado y guíado desde el día uno. Por la orientación, las revisiones y los comentarios que me permitieron hacer este trabajo. Por su paciencia y amabilidad en todo momento.

A la Mtra. Leticia Gracia Medrano, la Dra. Lizbeth Naranjo, la Mtra. Patricia Romero y al Dr. Raúl Rueda por sus invaluables observaciones tras la revisión de este trabajo.

A todos los profesores de los cursos que recibí, por haber dejado en mí mucho más que conocimientos matemáticos.

Al Dr. Jesús Eladio Sánchez y al Dr. Raúl Rueda, por haber sido un apoyo constante a lo largo de estos dos años.

A Lucy, María Inés y María Teresa, no solo por el apoyo brindado en todo momento, sino también por la preocupación y dedicación que tienen hacia cada uno de nosotros.

A Brenda, Corina, Darío y Edgar, por haber hecho de los cursos más entendibles, las horas de estudio más amenas y las tareas más cortas.

A Leticia, Ana y Luis Fernando, por demostrarme que se puede estar en familia aún estando lejos de casa.

A mis amigas, por estar ahí $24 \times 365 \times 14$.

A mi familia, porque fueron mi primera escuela e hicieron de mí quien soy hoy. Por haberme acompañado y apoyado siempre.

A todos aquellos que hicieron posible este final, mencionados o no,
mi más sincero agradecimiento...

Brenda

Índice

1. Introducción. Análisis de sobrevivencia **1**
1.1. Introducción 1
1.2. Análisis de sobrevivencia 2
1.2.1. Conceptos básicos 3
1.2.2. Métodos de análisis de sobrevivencia 11

2. Eventos Recurrentes **23**
2.1. Análisis de Eventos Recurrrentes 23
2.1.1. Deficiencias de los métodos tradicionales para el Análisis de Eventos Recurrentes 24
2.2. Métodos de Análisis de Eventos Recurrentes 26
2.2.1. Métodos basados en Modelos Lineales Generalizados 26
2.2.2. Métodos no basados en Modelos Lineales Generalizados 37

3. Aplicaciones **51**
3.1. Simulación 51
3.1.1. Datos simulados 51
3.1.2. Resultados 57
3.2. Datos reales 60
3.2.1. Problemática y datos 60
3.2.2. Resultados 61

4. Conclusiones **72**

Capítulo 1

Introducción. Análisis de sobrevivencia

En esta sección introduciremos el tema de modelos de eventos recurrentes y el problema específico que estudiaremos. También, con el objetivo de sentar las bases para los modelos de eventos recurrentes que se verán, haremos una breve revisión de los conceptos y métodos básicos de análisis de sobrevivencia.

1.1. Introducción

El análisis de sobrevivencia es un área estadística la cual se centra en el estudio del tiempo transcurrido desde un evento inicial, que determina la inclusión del individuo en el estudio, hasta un evento final, genéricamente llamado falla, el cual ocurre cuando el individuo presenta la característica para terminar el estudio. Es común en este tipo de estudio que el evento de interés no ocurra en todos los individuos o que el individuo abandone el estudio antes de que este le ocurra, situaciones que traerían como consecuencia que se registre solo información parcial sobre la variable de interés. El objetivo principal del análisis de sobrevivencia es incorporar esta información parcial que proporcionan los individuos censurados mediante métodos desarrollados para ese fin.

Existen eventos llamados cuales son repetibles; es decir, pueden ocurrirle dos o más veces a la misma unidad de observación. Ejemplos de estos eventos son los accidentes de tránsito, la reaparición de ciertos tipos de tumores, los tiempos de duración de ciertas máquinas luego de ser sometidas a reparaciones, entre otros. Si lo que se desea es observar y registrar varias apariciones del mismo evento, indudablemente es necesario implementar algunos métodos especializados de análisis de supervivencia que manejen estos datos de eventos recurrentes de forma adecuada y óptima. Una característica definitoria de este tipo de eventos es que las ocurrencias no pueden producirse de forma simultánea en un mismo sujeto lo que implica que existe una ordenación lógica de las mismas. Esta característica los diferencia de la ocurrencia de múltiples eventos, los cuales sí pueden coincidir en el tiempo sobre un mismo individuo.

Las ideas anteriores surgen a inicios de 1980 y dan paso a los modelos para eventos recurrentes, una nueva subcategoría de los métodos de análisis de sobrevivencia. A partir de este

momento se publican varias investigaciones donde se proponen diferentes métodos, cuyos objetivos frecuentemente implican, en primera medida, comprender y describir los procesos de eventos individuales, identificar y caracterizar la variación a través de una muestra de procesos, comparar grupos de procesos y por último, determinar la relación entre los predictores relevantes y la tasa en la que los eventos están ocurriendo.

En el presente trabajo no solo veremos los aspectos teóricos de estos métodos, sino también los aplicaremos a un conjunto de datos reales. Los datos que se utilizarán han sido extraídos de las bases de datos correspondientes a un ensayo clínico multicéntrico realizado por la Federación Francófona de Cancerología Digestiva 2000-05. Estos datos pertenecen a un seguimiento de pacientes con cáncer colorrectal metastásico e incluyen los tiempos de aparición de nuevas lesiones y muerte. Nuestro objetivo será describir este fenómeno y analizar su comportamiento desde el punto de vista de los eventos recurrentes. Por último, presentaremos una simulación que compara varios de los métodos que describiremos.

1.2. Análisis de sobrevivencia

Originalmente, el análisis de sobrevivencia se utilizó para investigaciones de mortalidad y morbilidad en estadísticas de registros vitales, siendo el análisis de procesos de sobrevivencia humana más antiguo que se conoce el realizado por el estadístico inglés John Graunt cuando publicó la primera tabla de vida en 1662. Durante un largo período de tiempo, este tipo de análisis se consideró un instrumento analítico, particularmente en estudios biomédicos y demográficos [18]. En una etapa posterior, se expandió gradualmente al dominio de la ingeniería para describir y evaluar el curso de los productos industriales, dando inicio así al análisis de la sobrevivencia tal como lo conocemos hoy.

En los últimos cuarenta años, el alcance del análisis de sobrevivencia ha crecido enormemente como consecuencia de los rápidos avances en la informática, en particular el avance de potentes paquetes de software estadísticos. Es así como ha llegado a tener un auge también en ciencias de la salud a partir de los años setenta. La ventaja que ofrecen estas técnicas y que en gran medida han contribuido a su popularización es que permiten generalizar el análisis de respuestas binarias (si/no; fallecido/vivo), incluyendo el tiempo de seguimiento, es decir, el tiempo que ha transcurrido desde el inicio del seguimiento hasta producirse la respuesta o hasta el final del seguimiento si la respuesta no se ha producido. Además, este tiempo que se analiza se puede valorar en condiciones muy flexibles, ya que la duración del período de observación puede ser muy diferente para cada sujeto.

Veremos a continuación una pequeña síntesis de este análisis, sus características y objetivos principales.

1.2.1. Conceptos básicos

Existen numerosos estudios longitudinales o de seguimiento donde la variable respuesta es binaria y los cuales se caracterizan por:

1. Duración variable del seguimiento: el seguimiento dado a cada individuo no dura lo mismo, ya que a pesar de que las fechas de inicio y de cierre del estudio están bien definidas, los sujetos se pueden incorporar al estudio en momentos diferentes.

2. Se tienen observaciones incompletas que dan lugar a datos censurados: estas se deben a varias razones, por ejemplo, puede suceder que en el estudio ciertos sujetos no presenten el evento esperado (sujetos retirados "vivos") o que se pierda el seguimiento de ciertos individuos (sujetos perdidos).

En la literatura nos encontramos con situaciones donde estos tipos de datos son analizados con varias técnicas estadísticas como regresión y pruebas χ^2, siendo una de las técnicas más utilizadas la regresión logística.

La *regresión logística* estudia una variable de respuesta binaria, generalmente codificada con ceros y unos, la cual registra la ocurrencia o no de un evento en particular en un periodo de tiempo fijo, o sea, no tiene en cuenta el momento exacto en el que ha ocurrido el evento. Sean las variables explicativas $\mathbf{x}^T = (x_1,\ x_2,\ ...,\ x_k)$ y denotando por $P(y = 1|\mathbf{x}) = \pi(\mathbf{x})$ la probabilidad condicional de que la variable de respuesta tome el valor 1 viene dada por

$$g(\mathbf{x_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

para $i = 1, 2, ..., n$ observaciones donde $g(x) = \log\left(\dfrac{\pi(x)}{1-\pi(x)}\right)$.

Tendríamos entonces que el modelo de regresión logística es:

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}.$$

En el caso en que una o varias variables explicativas de nuestro problema sean de tipo categóricas y tengan p niveles se generan $p-1$ variables ficticias (generalmente llamadas *dummys* o indicadoras) las cuales nos permiten representar correctamente todas las posibilidades de estas variables en el modelo utilizado. Si lo anterior sucede, por ejemplo para la variable explicativa j-ésima tendríamos que:

$$g(\mathbf{x_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \sum_{l=1}^{p-1} \beta_{jl} D_{jl} + ... + \beta_k x_{ik}.$$

El ajuste del modelo anterior se realiza mediante el método de máxima verosimilitud la cual se obtiene mediante $\widehat{\beta_q}$ para $q = 1, 2, ..., k$ y los vectores $\mathbf{x_i}$. A su vez, la estimación de los parámetros $\widehat{\beta_q}$ se obtiene mediante métodos iterativos como el método de Newton-Raphson.

En estos modelos el valor de un coeficiente β_q significa el incremento o disminución que se produce en el cociente

$$\frac{P(Y=1)}{P(Y=0)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}.$$

Un coeficiente positivo aumenta la probabilidad de ocurrencia del evento, ya que si β es positivo su transformación antilogaritmo será mayor que uno y la razón de probabilidades aumentará. Dicho aumento se produce cuando la probabilidad de ocurrencia $(Y = 1)$ de un suceso aumenta y la probabilidad de no ocurrencia de dicho evento $(Y = 0)$ disminuye, lo que trae consigo que según el modelo el evento tenga una elevada probabilidad de ocurrencia. Análogamente se interpreta el caso en el que el coeficiente estimado sea negativo. Por último, si tenemos algún caso donde el valor estimado del coeficiente sea 0, su antilogaritmo será 1, lo cual indicaría que no se produce ningún cambio en la razón de probabilidades.

A pesar de las diferentes técnicas estadísticas utilizadas para analizar este tipo de datos, el análisis de sobrevivencia es una de las mejores opciones para hacerlo ya que este incluye toda la información que aportan los datos y utilizan el tiempo como una variable explicativa, no así las anteriores mencionadas.

El *análisis de sobrevivencia* consiste en un tipo de técnica estadística especial la cual permite estudiar en particular una variable que se mueve a través del tiempo hasta que ocurre un evento de interés en la investigación [18]. Este análisis también incluye a las variables explicativas que proporcionan causalidad al desenlace de este evento. Otra característica importante de este es que permite que no se pierda una información muy importante: el ritmo al que se va presentando el evento, es decir, la tasa de incidencia del mismo.

Un **evento** se define como un cambio de estado en las condiciones iniciales del fenómeno que se está estudiando. Este puede situarse en algún punto del tiempo, es decir, una transición de un estado discreto inicial a otro. Algunos investigadores también definen un evento como el hecho de que una variable cuantitativa cruce un umbral establecido con anterioridad. Una condición necesaria para implementar el análisis de sobrevivencia, es identificar cuándo se produce un cambio de estado, es decir, ser capaz de situar el acontecimiento en el tiempo. También necesitamos un par de valores: el tiempo de seguimiento del sujeto y una variable binaria que indica si es un tiempo completo o censurado.

Los objetivos básicos del análisis de sobrevivencia son:

- Estimar e interpretar las funciones de sobrevivencia y/o riesgo de los datos.

- Comparar las funciones de sobrevivencia y/o riesgo.

- Evaluar la relación entre las variables explicativas y el tiempo de sobrevivencia.

Es posible realizar análisis de sobrevivencia cuando los datos constan únicamente de los tiempos de los acontecimientos, sin embargo un objetivo común del análisis de sobrevivencia es estimar modelos predictivos o causales en los que el riesgo de un evento depende de un conjunto de covariables. Si este es el objetivo de estudio, el conjunto de datos debe contener

las mediciones de las variables explicatorias en juego. Algunas de estas variables, tales como la raza y el sexo, pueden ser constantes a lo largo del tiempo. Otras, como los ingresos, el estado civil o la presión arterial, pueden variar con el tiempo. Por covariables de tiempo se debe entender aquellas variables con un gran conjunto de atributos que proporcionan tantos detalles del fenómeno como sea posible además de explicar la variación del mismo.

Algunos conceptos básicos involucrados en análisis de sobrevivencia son los siguientes [14]:

- **Fecha de inicio y fecha de cierre del estudio**: Determinan la duración del estudio.
- **Fecha inicial para cada sujeto**: Puede ser, por ejemplo, la fecha de diagnóstico o de inicio del tratamiento en este sujeto. Debemos tener en cuenta que los sujetos pueden entrar cada uno con fechas diferentes y que no tiene por qué coincidir estas con la fecha de inicio del estudio.
- **Fecha de última observación para cada sujeto**: La última noticia que se tiene de un sujeto marca la fecha de la última observación. Esta también marca el estado del sujeto y no tiene por qué coincidir con la fecha de cierre del estudio.
- **Seguimiento**: Observación de los individuos de un grupo a partir de la fecha inicial con el objetivo de conocer su estado.
- **Período de seguimiento**: Tiempo transcurrido entre la fecha de inicio y la fecha de cierre del estudio.
- **Tiempo de sobrevivencia**: Intervalo de tiempo transcurrido entre el inicio y la fecha de última observación.
- **Sujetos "retirados vivos"**: Se conoce así a los sujetos que se han seguido regularmente y que en el momento del cierre del estudio no han presentado el evento de interés.
- **Sujetos "perdidos"**: Se denominan de esta forma a los sujetos de los cuales se ha perdido el seguimiento.

Los conceptos básicos anteriores los podemos ver representados gráficamente en la Figura 1.1

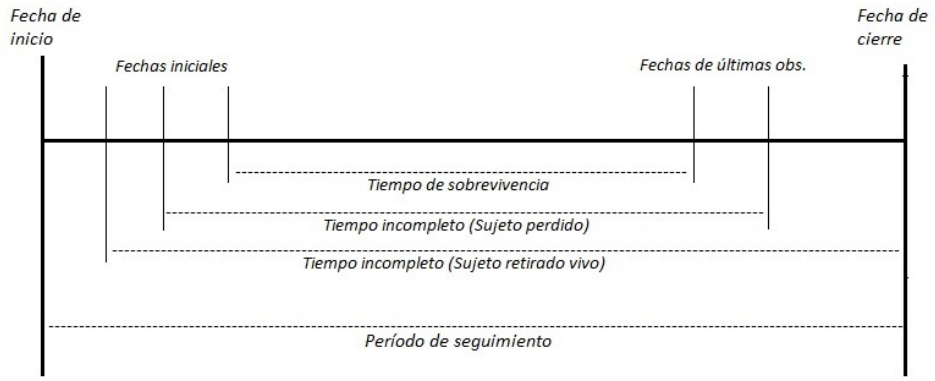

Figura 1.1: Representación gráfica de algunos conceptos básicos de análisis de sobrevivencia

El análisis de sobrevivencia tiene una característica que lo diferencia claramente de otros análisis estadísticos, y es que algunos individuos experimentan el evento terminal y otros no, lo que hace que el tiempo de supervivencia de los que no lo han experimentado sea un tiempo desconocido. Esta característica se define como **tiempo censurado** y puede tener varios orígenes:

a. el paciente no ha sufrido (aún) el evento terminal en la fecha de fin del estudio

b. se ha perdido el seguimiento del paciente

c. el paciente experimenta un evento diferente pero que imposibilita el seguimiento

Todas las anteriores dan lugar a tiempos censurados por la derecha los cuales son muy diferentes en origen pero son analizados de manera idéntica. Tendremos entonces **censura por la derecha** cuando para un individuo el tiempo T hasta el evento de interés no es observado pero sí se sabe que este es mayor a un tiempo s donde corresponde al último tiempo de seguimiento. En general, para una muestra de n individuos, los datos vendrían representados por $\{(Y_1, \delta_1), (Y_2, \delta_2), ..., (Y_n, \delta_n)\}$ donde:

$$Y = \text{mín}\{T_i, s_i\}, \quad \delta_i = \begin{cases} 1 & \text{si } T_i \leq s_i : \text{el individuo i no está censurado.} \\ 0 & \text{si } T_i > s_i : \text{el individuo i está censurado.} \end{cases}$$

siendo T_i y s_i los valores correspondientes de T y s para el individuo i.

La censura por la derecha puede ser censura tipo I (fija, progresiva o generalizada) o censura tipo II.

Estaremos en presencia de censura **tipo I** cuando el evento se observa sólo si ocurre antes de un tiempo preespecificado. Esta se clasifica de acuerdo a lo siguiente:

- **fija:** si la censura es la misma para todos los individuos
- **progresiva:** si la censura es diferente para distintos grupos de individuos
- **generalizada:** si la censura es distinta para cada individuo

La censura **tipo II** se tiene cuando se decide finalizar el estudio después que hayan ocurrido un número predeterminado de m eventos, o sea, el estudio culmina cuando se observe el evento m-ésimo.

Notemos que estos tiempos censurados por la derecha están subestimando la verdadera sobrevivencia del sujeto, y por tanto, el estudio subestima la sobrevivencia de la población.

Es importante señalar que, a pesar de que generalmente los tiempos con los que trabajamos están censurados por la derecha, como es el caso de los retirados vivos y de los perdidos, también podemos encontrar situaciones donde los tiempos pueden ser censurados por la izquierda

o, incluso, podemos tener intervalos censurados.

Una observación se dice **censurada por la izquierda** si se desconoce el valor exacto del tiempo hasta el evento de interés pero se sabe que ha ocurrido antes de comenzar el seguimiento del individuo. Por su parte, tendremos **intervalos censurados** cuando desconocemos el tiempo hasta el evento de interés pero sabemos que se encuentra en cierto intervalo, por ejemplo, cuando un evento se produce entre dos consultas médicas sabemos que se ha producido entre dos tiempos, pero no sabemos el valor exacto del tiempo en que ocurrió.

Sobrevivencia, distribución, riesgo y riesgo acumulado

Existen diferentes funciones de los datos de sobrevivencia las cuales nos aportan información importante, y por tanto es útil su estudio. Analicemos a grandes rasgos cada algunas de ellas.

La *función de sobrevivencia*, también llamada *probabilidad de sobrevivencia* y denotada por $S(t)$, es la probabilidad de que un individuo sobreviva desde la fecha de entrada en el estudio hasta un momento determinado en el tiempo t [18], o sea:

$$S(t) = \mathbb{P}(T > t).$$

La *función de distribución* representa la probabilidad de que a un individuo le ocurra el evento antes del tiempo t para $t \geq 0$ y se define como :

$$F(t) = \mathbb{P}(T \leq t).$$

Estas dos funciones están relacionadas ya que $F(t) = 1 - S(t)$.

Si T es continua, con función de densidad $f(t)$, entonces tendremos que:

$$S(t) = \mathbb{P}(T > t) = \int_t^\infty f(u)\mathrm{du} \quad \text{y} \quad F(t) = \mathbb{P}(T \leq t) = \int_0^t f(u)\mathrm{du}.$$

Si T es discreta con función de masa de probabilidad $f(t_j) = \mathbb{P}(T = t_j)$, $j = 1, 2, ...$ y toma los valores $t_1 < t_2 < ...$, su función de sobrevivencia y de distribución vienen dadas por:

$$S(t) = \mathbb{P}(T > t) = \sum_{t_j > t} f(t_j) \quad \text{y} \quad F(t) = \mathbb{P}(T \leq t) = \sum_{t_j \leq t} f(t_j).$$

Todas las funciones de sobrevivencia tienen las siguientes propiedades teóricas:

- Son no crecientes, o sea, se mantienen constantes o disminuyen a medida que t aumenta.
- En el tiempo $t = 0$ se tiene que $S(t) = S(0) = 1$, o sea, al comienzo del estudio como nadie ha tenido el evento aún la probabilidad de sobrevivir al tiempo 0 es 1 .
- En el tiempo $t = \infty$ se tiene que $S(t) = S(\infty) = 0$, o sea, si el período de estudio se incrementa sin límites, eventualmente nadie sobrevivirá por lo que en algún momento la curva de sobrevivencia caerá a 0.

- Como t va desde 0 hasta ∞, la función se puede graficar como una curva suave.

Debemos tener en cuenta que estas son propiedades teóricas de curvas de sobrevivencia, en la práctica, cuando se utilizan datos reales se suelen obtener gráficos que son funciones escalonadas en lugar de curvas suaves. También, debido a que el período de estudio no es infinito, puede haber riesgos competitivos para el fracaso y es posible que no todos los individuos del estudio presenten el evento estudiado [15]. En la Figura 1.2 mostramos un ejemplo de una gráfica de la función de sobrevivencia teórica y de la práctica.

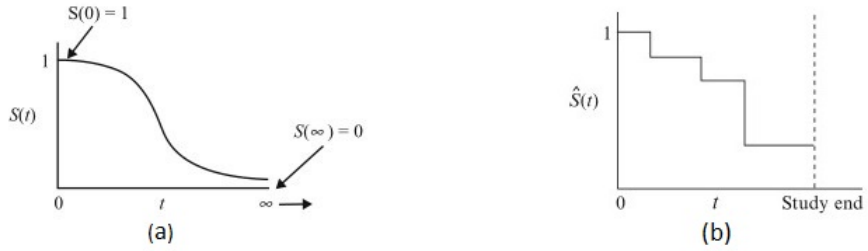

Figura 1.2: a-Función de sobrevivencia teórica. b-Función de sobrevivencia real

La *función de riesgo*, $h(t)$, a pesar de ser menos conocida que la función de sobrevivencia, en algunos casos resulta de mayor interés. Esta denota la "probabilidad" de que a un individuo que está siendo observado en el tiempo t le suceda el evento en ese momento [18]. En algunos casos también la vemos interpretada como que nos da el potencial instantáneo en el tiempo t para obtener un evento, dada la sobrevivencia hasta ese tiempo.

Esta función, a veces también conocida como *tasa de falla condicional*, se define en el caso continuo como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

donde Δt denota un pequeño intervalo de tiempo.

En este caso esta función también puede expresarse como

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{dt}}(\ln S(t)).$$

Para el caso discreto donde la variable T toma los valores $t_1 < t_2 < ...$ y tiene función de masa de probabilidad $f(t_j) = \mathbb{P}(T = t_j)$ para $j = 1, 2, ...$ tendremos que

$$h(t_j) = \mathbb{P}(T = t_j | T \geq t_j), \quad j = 1, 2, ...$$

Se tienen también las siguientes expresiones que relacionan las funciones de sobrevivencia y de riesgo:

$$h(t_j) = \frac{f(t_j)}{S(t_{j-1})} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})} \quad \text{y} \quad S(t) = \prod_{j:t_j \leq t} (1 - h(t_j)).$$

Es importante notar que la función de riesgo para el caso de una variable continua define una tasa en lugar de una medida de probabilidad. Lo anterior se debe a que los valores para esta función no oscilan entre 0 y 1, sino entre 0 e ∞ y depende en gran medida de la unidad de tiempo usada. En particular, para un valor de t, la función de riesgo $h(t)$ se caracteriza por ser siempre no negativa y por no tener un límite superior. Al igual que la función de sobrevivencia esta función se puede representar gráficamente pero se diferencia de ésta en que no tiene que comenzar en 1 y decrecer a 0, sino que puede empezar en cualquier valor y crecer o decrecer en cualquier dirección en el tiempo.

Como hemos visto, la información obtenida mediante la función de sobrevivencia y la de riesgo son muy diferentes, ya que la primera se centra sobre todo en la "no ocurrencia" del evento mientras que la segunda se centra en la "ocurrencia" del evento y proporciona información tan valiosa como la tasa de incidencia del evento. Debido a lo anterior, en cierto sentido, se puede considerar que la función de riesgo nos da el lado opuesto de la información dada por la función de sobrevivencia.

De las funciones consideradas $S(t)$ y $h(t)$, la función de sobrevivencia es más común en el análisis de estos datos debido a que ella describe directamente la experiencia de sobrevivencia de una cohorte de estudio. Sin embargo, la función de riesgo también es de interés por las siguientes razones:

- Es una medida de potencial instantáneo, mientras que una curva de sobrevivencia es una medida acumulativa sobre el tiempo.

- Puede usarse para identificar la forma de un modelo específico, tal como exponencial, Weibull o una curva logarítmica normal que se ajuste a los datos.

- Es el vehículo que permite modelar de forma matemática los datos de sobrevivencia, es decir, el modelo de sobrevivencia generalmente se escribe en términos de la función de riesgo.

Independientemente de la función con la que se esté trabajando, ya sea $S(t)$ o $h(t)$, existe una relación claramente definida entre ellas, por lo que si conocemos la expresión de $S(t)$ podemos obtener la expresión de $h(t)$ y viceversa [15]. Esta relación entre las funciones se puede expresar de forma equivalente por las siguientes fórmulas generales

$$S(t) = \exp\left[-\int_0^t h(u)\ \mathtt{d}u\right] \quad \text{y} \quad h(t) = -\left[\frac{\mathtt{d}S(t)/\mathtt{d}t}{S(t)}\right]$$

.

Por último, veamos la *función de riesgo acumulado* $\Lambda(t)$, la cual a pesar de no tener una interpretación intuitiva clara, suele ser muy útil desde un punto de vista técnico.

Si T una variable continua con función de riesgo $h(t)$ definiremos a la función de riesgo acumulado como

$$\Lambda(t) = \int_0^t h(u)\ \mathrm{du}.$$

Utilizando la expresión de la función de sobrevivencia tendremos que:

$$S(t) = \exp\left[-\Lambda(t)\right] \quad \text{y} \quad \Lambda(t) = -\ln S(t).$$

Si T es una variable discreta la cual toma los valores $t_1 < t_2 < ...$ y cuya función de riesgo es $h(t_j)$, definiremos la función de riesgo acumulado como

$$\Lambda(t) = \sum_{j:t_j \leq t} h(t_j).$$

En este caso ya no tendremos la relación entre la función de riesgo acumulado y la función de sobrevivencia como en el caso anterior.

Queremos destacar que dependiendo de la monotonía y del comportamiento de las funciones de riesgo, se pueden clasificar los modelos de sobrevivencia. Entre estas clasificaciones encontramos las siguientes:

Modelo de sobrevivencia exponencial: Cuando la función de riesgo es constante, o sea, $h(t)$ toma el mismo valor para todo t (Figura 1.3:a).

En este caso tendríamos:

$$h(t) = \theta \in \mathbb{R} \;\; \text{y} \;\; S(t) = \exp\{-\theta t\}.$$

Modelo de sobrevivencia Weibull: Cuando la función de riesgo mantiene una monotonía constante en el tiempo. En dependencia de si aumenta o disminuye el modelo se conoce como creciente o decreciente respectivamente. (Figura 1.3:b)

Para este tenemos que $T \sim \text{Weibull}(\lambda, \gamma)$ si $T^\gamma \sim \text{Exp}(\lambda)$ por lo que

$$h(t) = \gamma\lambda^p t^{\gamma-1} \;\; \text{y} \;\; S(t) = \exp\{-(\lambda t)^\gamma\}.$$

Los dos modelos anteriores los analizaremos con más detalle posteriormente.

Modelo de sobrevivencia lognormal: Cuando la función de riesgo primero aumenta y luego disminuye. (Figura 1.3:c)

En este caso tenemos que T tiene una distribución lognormal si $Y = \log T = \alpha + \sigma W$ donde W tiene una distribución normal estándar.

Como mencionamos anteriormente, en este caso la función de riesgo de la distribución log-normal crece desde 0 hasta alcanzar su máximo y después decrece monótonamente aproximándose a 0 a medida que $t \to \infty$

Tendremos que la función de densidad de una distribución gamma generalizada, la cual se escribe como:

$$f(t) = \frac{\lambda p(\lambda t)^{pk-1}\mathsf{e}^{-(\lambda t)^p}}{\Gamma(k)}$$

con $p = 1/\sigma$ se aproxima a una lognormal para el caso límite cuando $k \to \infty$.

Notemos que de esta distribución se pueden obtener como casos particulares las dos anteriores, exponencial cuando $p = 1$ y $k = 1$ y Weibull cuando $k = 1$.

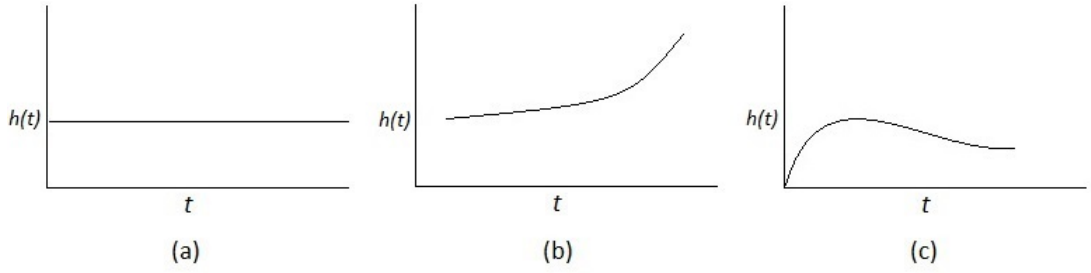

Figura 1.3: Modelos de sobrevivencia: a-exponencial, b-Weibull, c-lognormal

1.2.2. Métodos de análisis de sobrevivencia

Como vimos anteriormente, el análisis de sobrevivencia tiene dos características comunes que son difíciles de manejar con los métodos estadísticos convencionales: la censura y las covariables que dependen del tiempo, comúnmente llamadas variables explicativas. Al contrario de los métodos convencionales, los cuales no ofrecen mucha esperanza para tratar con cualquier censura o covariables dependientes del tiempo, todos los *métodos de análisis de sobrevivencia* se adaptan para lidiar con estas características de diferentes formas.

Las técnicas disponibles para el análisis de sobrevivencia pueden agruparse en dos grandes grupos [12]:

- las técnicas descriptivas e inferenciales univariadas, como la técnica de Kaplan-Meier y las comparaciones de LogRank.

- las técnicas inferenciales multivariadas de riesgos proporcionales que ajustan diversos modelos estadísticos.

Estas últimas pueden dividirse a su vez en métodos semiparamétricos, los cuales consideran algunos supuestos sobre los datos; y métodos paramétricos, los cuales ajustan una distribución de probabilidad en los datos. Ambos tipos de métodos permiten incluir covariables, ya sean independientes o dependientes del tiempo, lo que facilita ver la influencia de otros factores en el tiempo hasta que sucede un evento.

A continuación resumiremos brevemente algunos métodos que pertenecen a cada grupo.

Técnicas descriptivas no parámetricas

Estos modelos se caracterizan por estimar la función de sobrevivencia sin tomar en cuenta la distribución de los datos, únicamente considerando la información provista por la muestra.

Una de las técnicas más conocidas y utilizadas es el **estadístico de Kaplan-Meier**, también conocido como **estimador producto-límite**. Este fue introducido por Edward Kaplan y

Paul Meier en [13]. En él se tienen en cuenta el número de eventos que suceden en un determinado tiempo y el número de sujetos en riesgo antes del tiempo, tomando en consideración también la existencia de censura en los datos. Dado que se asume que los eventos ocurren de forma independiente el uno del otro, las probabilidades de sobrevivencia de un tiempo hasta otro tiempo son multiplicadas siguiendo el principio multiplicativo de la probabilidad, de forma que obtenemos la función de sobrevivencia.

Formalmente, este estimador viene dado por:

$$\widehat{S}(t) = \begin{cases} 1 & \text{si } t < Y(1) \\ \prod_{t_i \leq t}(1 - \widehat{q}_i) = \prod_{t_i \leq t}\left(1 - \dfrac{d_i}{n_i}\right) & \text{si } t \geq Y(1), \end{cases}$$

donde n_i es el número de individuos en riesgo al tiempo t_i, d_i el número de individuos que "fallaron" hasta este tiempo y $\widehat{q}_i = d_i/n_i$ es la probabilidad de sobrevivencia condicional.

La función estimada a partir de este método puede ser representada gráficamente donde, generalmente, el eje X representa el tiempo desde que inicia el estudio hasta que finaliza el seguimiento y el eje Y la probabilidad de que un individuo no presente el evento. A estos se les conocen como *gráficos de sobrevivencia* o *curva de sobrevivencia* y su interpretación suele ser bastante sencilla. A pesar de que no es lo más común también se puede trabajar con la función de riesgo para realizar gráficos similares.

Una de las ventajas que tiene este método es que permite la estimación de las funciones de sobrevivencia de diferentes grupos y posteriormente se pueden utilizar medidas estadísticas para determinar si en dos grupos o más son iguales o diferentes estas funciones [12]. Lo anterior es posible gracias a la adaptación de estadísticos no paramétricos, por ejemplo, el **estadístico LogRank** el cual constituye una adaptación del estadístico de Mantel-Haenszel; y los **estadísticos generalizados de Wilcoxon** que no son más que versiones generalizadas de la suma de rangos de Wilcoxon . Veamos algunas de las características de estos estadísticos.

Frecuentemente en los estudios de sobrevivencia cuando se comparan las experiencias de dos o más grupos de pacientes encontramos que estos difieren con respecto a cierto factor. El efecto de este factor en la sobrevivencia es el evento de interés al comparar los grupos y para detectarlo se suele realizar la prueba LogRank.

La prueba LogRank, como mencionamos anteriormente, es un método estadístico no paramétrico el cual está diseñado para detectar si existen o no diferencias entre las curvas de sobrevivencia mediante la comparación de la experiencia de sobrevivencia de dos o más grupos. Estas diferencias ocurren cuando la tasa de mortalidad en un grupo es consistentemente mayor que la tasa correspondiente a los otros grupos con los que se esté comparando. A grandes rasgos, esta prueba consiste en una prueba de hipótesis donde la hipótesis nula (H_0) es aquella que postula que no existen diferencias en la experiencia de sobrevivencia de los grupos. Para realizarla es necesario calcular el número de eventos observados y esperados para cada periodo de tiempo, los cuales son sumados individualmente. También para cada periodo de tiempo en

el que ocurre un evento, independientemente del grupo en el que ocurran, se calcula el número de pacientes en riesgo y el número de eventos observados para cada grupo. Utilizando estos datos se calcula el número esperado de eventos asumiendo la veracidad de la hipótesis nula. Las ecuaciones para calcular el número de eventos esperados (e_{ij}), cada vez que ocurre un evento en cada grupo, son [1]

$$e_{ij} = \frac{r_{ij} d_i}{r_i} \quad , \quad d_i = \sum_{j=1}^{g} d_{ij} \quad \text{y} \quad r_i = \sum_{j=1}^{g} r_{ij},$$

en las cuales:

e_{ij}: número de eventos esperados en el i-ésimo intervalo de tiempo para el j-ésimo grupo bajo estudio,

r_{ij}: número de pacientes en riesgo de fallecer en el i-ésimo intervalo de tiempo y quienes pertenecen al j-ésimo grupo bajo estudio,

d_i: número total de eventos observadas en el i-ésimo intervalo de tiempo del periodo bajo estudio,

d_{ij}: número de eventos observados en el i-ésimo intervalo de tiempo para el j-ésimo grupo bajo estudio,

r_i: número total de pacientes en riesgo de presentar el evento en el i-ésimo intervalo de tiempo del periodo bajo estudio, y

g: número total de grupos.

Los valores de los eventos observados y esperados son a su vez sumados para obtener los siguientes:

$$O_j = \sum_{i=1}^{t} d_{ij} \qquad E_j = \sum_{i=1}^{t} e_{ij}$$

donde:

O_j: número total de eventos observados durante la duración del estudio para el j-ésimo grupo,

E_j: número total de eventos esperados durante la duración del estudio para el j-ésimo grupo,

i: tiempo en el que ocurre un evento, y

t: número total de tiempos cuando ocurren los eventos.

Luego de obtener el número total de eventos, tanto observados como esperados, se calcula la estadística LogRank [1]

$$U_L = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{E_j},$$

en la cual U_L tiene una distribución chi-cuadrada con $g - 1$ grados de libertad ($\chi^2_{(g-1)}$).

Debemos tener en cuenta que al ser esta, simplemente una prueba de hipótesis ella no provee información directa de cómo los grupos difieren. Para medir la sobrevivencia relativa entre dos grupos se deben comparar los eventos observados con los esperados hallando la razón entre ellos para cada grupo. Al realizar lo anterior, obtendremos la tasa observada de ocurrencia de eventos en ese grupo como una proporción de la tasa esperada de ocurrencia de eventos si la hipótesis nula de la prueba LogRank fuera cierta. Utilizando esta razón entre dichas tasas de ocurrencia, se puede calcular la experiencia relativa de sobrevivencia de dos grupos, conocida también como la razón de riesgo.

Por ejemplo, sea O_1/E_1 (O_2/E_2) la razón entre las tasas de ocurrencia para los individuos que pertenecen al grupo 1 (grupo 2) la razón de riesgo entre estos dos grupos y su intervalo de confianza vendrán dados por las siguientes expresiones [1]:

$$\text{RR} = \frac{O_1/E_1}{O_2/E_2} \quad \text{y} \quad \text{IC} = \Big(\exp[b - Z_{1-\alpha/2}(1/\sqrt{V})],\ \exp[b + Z_{1-\alpha/2}(1/\sqrt{V})]\Big)$$

donde

$$b = \frac{(O_1 - E_1)}{V} \quad \text{y} \quad V = \sum_{i=1}^{t} v_i = \sum_{i=1}^{t} \frac{r_{i1} r_{i2} d_i (r_i - d_i)}{{r_i}^2 (r_i - 1)}$$

en las cuales

$Z_{1-\alpha/2}$: cuantil $1 - \alpha/2$ de la distribución normal estándar,

b: estimado del logaritmo de la razón de riesgo,

O_1 (O_2): tasa observada de ocurrencia en el grupo 1 (2),

E_1 (E_2): tasa esperada de ocurrencia en el grupo 1 (2),

V: suma de las varianzas v_i,

v_i: varianzas entre los eventos observados y los eventos esperados en el i-ésimo tiempo de sobrevivencia, y

r_{i1} (r_{i2}): número de pacientes en riesgo de presentar el evento en el i-ésimo tiempo de sobrevivencia y que pertenecen al grupo 1 (2).

Vale destacar que esta prueba es bastante robusta contra desviaciones de azar proporcional, o sea, desviaciones de los logaritmos de las curvas de sobrevivencia, no obstante debe tenerse cuidado ya que si las curvas de sobrevivencia de Kaplan-Meier se cruzan, esto sería evidencia de que los tiempos de sobrevivencia tienen mayor varianza en uno de los grupos y por tanto esta prueba no sería una técnica apropiada [4]. En caso de que suceda lo anterior se utilizan otras pruebas tales como las pruebas generalizadas de Wilcoxon.

La prueba de Wilcoxon generalizada de Gehan consiste en una generalización de la prueba de Wilcoxon para dos muestras para el caso en que se tienen datos censurados. Esta se basa en el estadístico:

$$U_W = \sum_{j=1}^{g} r_j (O_j - E_j)$$

cuya varianza es

$$V_W = \sum_{j=1}^{g} {r_j}^2 v_j.$$

Utilizando las anteriores se define el estadístico de prueba de la siguiente forma

$$E_W = U_W^2 / V_W.$$

El estadístico anterior tiene una distribución asintótica ji-cuadrada con $g-1$ grados de libertad bajo la hipótesis nula. Este difiere de la estadística de LogRank por los pesos r_j.

Otra prueba Wilcoxon generalizada, pero en este caso para observaciones no censuradas se describe en [20]. En esta utilizan como peso la estimación de Kaplan Meier de la función de sobrevivencia ($\widehat{S}(t)$)

$$U_P = \sum_{j=1}^{g} \widehat{S}(t)(O_j - E_j).$$

Esta prueba se reduce a la prueba de Wilcoxon de Gehan cuando no hay observaciones censuradas.

En general, la prueba LogRank tiende a ser sensible a las diferencias de distribución las cuales son más evidentes al final del tiempo; mientras que las pruebas generalizadas de Wilcoxon tienden a ser más sensibles para detectar diferencias temprano en el tiempo como consecuencia de la diferencia en los pesos que se otorgan a cada evento [21]. A pesar de que los p-valores que se obtienen de ambas pruebas generalemente son similares, está demostrado que cuando la razón de riesgo no es constante, las pruebas de Wilcoxon generalizadas puede ser más efectivas que la prueba LogRank [17]. Debido a lo anterior, en las aplicaciones la prueba LogRank se utiliza después de verificar la validez de la suposición de riesgos proporcionales, siendo las de Wilcoxon generalizadas el método alternativo cuando falla este supuesto. Como deficiencia de ambas pruebas podemos mencionar que ninguna nos provee de una medida general de asociación, como un riesgo relativo, y tampoco de intervalos de confianza.

A pesar de que las pruebas anteriores son las más conocidas y utilizadas para comparar las funciones de sobrevivencia de dos o maś grupos, debemos mencionar que existen otras pruebas como la de Tarone-Ware la cual está considerada como una prueba intermedia entre las dos pruebas analizadas anteriormente.

Antes de concluir con las pruebas utilizadas para comparar funciones de sobrevivencia de dos o más grupos queremos enfatizar en las características comunes de los tres tipos de pruebas mencionadas (incluyendo la de Tarone-Warner):

- **Hipótesis nula** (H_0): las funciones de sobrevivencia de los grupos son iguales.
- **Hipótesis alternativa** (H_1): al menos uno de los grupos tiene una función de sobrevivencia diferente.
- **Estadístico utilizado:** se distribuye como χ^2 con $g-1$ grados de libertad, siendo g el número de grupos que se comparan.

Retomando el tema de las técnicas para estimar las funciones de sobrevivencia de una muestra, debemos señalar que existen otros métodos a pesar de ser el método de Kaplan-Meier el más utilizado. Ejemplo de otro método sería el **estadístico de Nelson-Aalen**, el cual permite estimar esta función a partir de la función de riesgo de la muestra [12]. Este estimador para la función de riesgo acumulado se define de la siguiente forma

$$\widehat{\Lambda}_{NA}(t) = \begin{cases} 0 & \text{si } t < Y(1) \\ \sum_{i:t(i)\leq t} \frac{d_i}{n_i} & \text{si } t \geq Y(1), \end{cases}$$

donde d_i representa el número de fallos ocurridos en el instante t_i, n_i el número de individuos en riesgo en t_i y $Y(1)$ el menor tiempo. Una estimación de la función de riesgo en t_i a partir de este estimador se puede obtener como d_i/n_i.

Es importante destacar que pese a que los estimadores de Kaplan-Meier y de Nelson-Aalen presentan una eficiencia relativamente similar en la práctica, desde un punto de vista estadístico, la propuesta de Nelson y Aalen sienta las bases para uno de los métodos de análisis de sobrevivencia más utilizados, el cual se abordará a continuación.

Técnicas semiparámetricas que incluyen covariables

Las técnicas anteriormente explicadas son útiles en el sentido en que nos permiten identificar la función de sobrevivencia de uno o varios grupos, e incluso compararlos, pero no nos permiten conocer la magnitud de las diferencias de estos o considerar otras covariables para evaluar su efecto en el tiempo hasta que suceda el evento de interés. Una técnica que sí permite considerar esto son los modelos de regresión que toman en cuenta la existencia de datos censurados. Uno de los modelos más extendidos en su uso es el *modelo de regresión de Cox de riesgos proporcionales* [7].

Este modelo se clasifica como una técnica semiparamétrica ya que hace uso de la estimación no paramétrica propuesta por la técnica de Nelson-Aalen, permite la inclusión de covariables [12] y además requiere el cumplimiento del supuesto de riesgos proporcionales. Este supuesto significa que la función de riesgo de diferentes valores de una covariable es proporcional a lo largo del tiempo, es decir, el riesgo es independiente del tiempo. La verificación de este supuesto resulta esencial para poder considerar válidos los resultados y predicciones brindados por el modelo resultante. El supuesto de proporcionalidad se puede evaluar de forma gráfica de la siguiente manera:

- Verificando que no se crucen las líneas del gráfico de las curvas de sobrevivencia.
- Comprobando que sean aproximadamente paralelas las líneas del gráfico loglog: $\texttt{ln}[-\texttt{ln}(S)]$ `vs.` `tiempo` para todos los grupos.

Vale destacar que en caso de que el supuesto de riesgos proporcionales no se cumpla se debe utilizar como alternativa el modelo de Cox estratificado.

La ecuación del modelo de regresión de riesgos proporcionales de Cox es la siguiente

$$h(t;\mathbf{X}) = h_0(t)\ \exp\{\beta^t\mathbf{X}\} = h_0(t)\ \exp\{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\},$$

donde $h(t;\mathbf{X})$ representa el riesgo de fallecer en el instante t de los sujetos que tienen un determinado patrón de valores x en las variables explicativas.

Como podemos ver el modelo anterior está compuesto por el producto de dos términos, donde:

- $\exp\{\beta^t\mathbf{X}\}$: función exponencial cuyo exponente es la combinación lineal, sin término constante, de las p variables explicativas x_i.

- $h_0(t)$: función de riesgo de referencia (conocida en inglés como "baseline" o "underlyng hazard function "). Esta sólo depende del tiempo y representa las tasas instantáneas de riesgo de un sujeto hipotético con valor 0 en todas las variables predictivas.

El modelo anterior se caracteriza por no especificar la forma de $h_0(t)$. En este se estiman los p parámetros β_i mientras que el valor de la función $h_0(t)$ se obtiene a través de los datos.

Este modelo tiene cierta similitud con el modelo de regresión logística expresado en forma de razón de probabilidades de sujetos codificados con $Y = 1$, o sea, con "respuesta" presente.

$$\frac{\pi_x}{1-\pi_x} = \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\} = \exp\{\beta_0\}\ \exp\{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\}.$$

Podemos comprobar que el término constante $\exp\{\beta_0\}$ en la expresión anterior correspondería en el modelo de regresión de Cox a la función de riesgo de referencia $h_0(t)$, la cual, al igual que este término, no depende de las variables explicativas $\mathbf{X}$.

El modelo de Cox lo podemos expresar en función de la función de sobrevivencia $S(t)$ o de la función de riesgo acumulado $\Lambda(t)$ debido a la relación existente entre ambas funciones. Tendríamos entonces que para este modelo

$$\Lambda(t) = h_0(t)\exp\{\beta^t\mathbf{X}\} \quad \text{y} \quad \Lambda(t) = -\texttt{ln}S(t),$$

donde la última igualdad se tiene a través de la relación de la función de sobrevivencia con la función de riesgo acumulado, y entonces $S(t) = \exp\{-\Lambda(t)\}$.

Para obtener los valores estimados de los coeficientes de regresión β_i correspondientes a las variables predictoras x_i en el modelo de regresión de riesgos proporcionales de Cox, se utiliza el método de máxima verosimilitud. Como parte de este método se determina la función de verosimilitud L, la cual puede ser obtenida teóricamente a través de la función de verosimilitud parcial. Esta función elimina la función base de riesgo desconocida denotada por $h_0(t)$, además de que no es afectada por los tiempos de sobrevivencia censurados. La fórmula para obtener la verosimilitud del modelo sería

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp\{\sum_{i=1}^{p} \beta_i^t\ x_{ij}\}}{\sum_{m\in R_j} \exp\{\sum_{i=1}^{p} \beta_i^t\ x_{im}\}}$$

donde:

- β_i: coeficiente de regresión correspondiente a la variable predictora x_i,

- x_i: el valor de la i-ésima variable predictora,

- R_j: conjunto de individuos a riesgo que todavía están vivos justo antes del tiempo t_j,

- t_j: j-ésimo tiempo conocido en el que ocurre un evento,

- k: número total de tiempos en los que ocurre un evento,

- p: número total de variables predictoras y

- m: individuos del conjunto R_j a riesgo en el tiempo t_j.

Sobre la verosimilitud anterior consideramos importante hacer notar que no es una verosimilitud en el sentido usual ya que tiene algunas características particulares, por ejemplo:

- El número de términos considerados no es el número de individuos en la muestra, sino el de fallas.
- Si observamos un individuo más o quitamos un individuo de la muestra podemos provocar que cambien los conjuntos en riesgo y, por tanto, la nueva verosimilitud no sería proporcional a la anterior.
- La verosimilitud depende únicamente del rango de los tiempos de fallas ya que estos determinan el conjunto en riesgo de cada tiempo de falla.

Para hallar los estimadores de β_i debemos derivar la logverosimilitud e igualarla a cero. Generalmente las ecuaciones obtenidas al realizar lo anterior se resuelven por métodos númericos como el Método de Newton-Raphson.

El procedimiento anterior se lleva a cabo en el caso en que los datos no contengan tiempos observados empatados, ya que cuando sí los contienen la verosimilitud parcial puede no ser inmediata y generalmente lleva un tiempo de computación considerable obtenerla. Debido a esto, en presencia de datos con empates, se suelen utilizar aproximaciones para la función de verosimilitud parcial. En este caso existen varias propuestas para construir dichas aproximaciones entre las que destacan las propuestas por Breslow (1974) y Efron (1982), esta última considerada más exacta que la primera. Las verosimilitudes con las que se trabajan las detallamos a continuación.

Supongamos J tiempos de falla para n individuos no censurados y $n - J$ censurados a derecha. Sean $t_{(1)} < t_{(2)} < ... < t_{(J)}$ los tiempos ordenados distintos con sus correspondientes covariables $x_1,\ x_2,\ ...,\ x_p$. Considerando que los d_j eventos al tiempo t_j son distintos y ocurren secuencialmente. Las aproximaciones vienen dadas por las siguientes fórmulas:

- Breslow:

$$L(\beta) = \prod_{j=1}^{J} \frac{\exp\{\beta^t\ S_j\}}{\left[\sum_{m \in R_j} \exp\{\beta^t\ X_m\}\right]^{d_j}},$$

donde S_j es la suma de los vectores X_p sobre todos los individuos que presentan el evento en t_j.

- Efron:

$$L(\beta) = \prod_{j=1}^{J} \frac{\exp\{\beta^t\ X_j\}}{\prod_{i=1}^{d_j} \left[\sum_{m\in R_j} \exp\{\beta^t\ X_m\} - \frac{i-1}{d_j} \sum_{m\in D_i} \exp\{\beta^t\ X_m\}\right]},$$

donde D_i representa el conjunto de etiquetas de los individuos que presentan el evento en el tiempo t.

Para concluir con el tema de estimación de los coeficientes del modelo debemos mencionar también que en algunas ocasiones se emplean para realizar esta tarea métodos de remuestreo. Estos consisten en una técnica de simulación basada en la información de interés la cual reduce la necesidad de asumir determinado modelo probabilístico para las observaciones. Lo anterior resulta una gran ventaja, no obstante, estos tienen también desventajas como por ejemplo que generalmente requieren un número elevado de cálculos. Algunos métodos de remuestreo empleados son el Método Bootstrap y el Método Jackknife.

Para determinar cuáles variables explicativas son significativas se determinan varios modelos de riesgos proporcionales que contengan diferentes combinaciones de dichas variables. Estos modelos son comparados entre sí para determinar cuál de todos explica mejor la situación de sobrevivencia en los datos estudiados. Para esto, se pueden utilizar los estadísticos de devianza, el criterio de información de Akaike (AIC) y el criterio de información Bayesiano (BIC), todos basados en el logaritmo de la verosimilitud. Se debe tener en cuenta que tanto el AIC como el BIC tienen una penalización basada en el número de parámetros y este último también está penalizado dependiendo del tamaño de la muestra; aunque en el caso de datos de sobrevivencia se recomienda utilizar en lugar del tamaño de la muestra el número de eventos observados.

Mediante el modelo de riesgos proporcionales de Cox puede estimarse la función de sobrevivencia de una muestra utilizando el diagnóstico brindado por una prueba como una covariable para controlar su efecto en el tiempo hasta que ocurre el evento. Esto permite obtener la llamada "*razón de riesgo*" *(risk ratio)*, la cual se define en general como la razón entre el riesgo para un conjunto de valores de las variables explicativas ($\mathbf{X^1}$) y el riesgo para otro conjunto de valores de estas variables ($\mathbf{X^2}$), o sea:

$$\widehat{HR} = \frac{h(t, \mathbf{X^1})}{h(t, \mathbf{X^2})} = \frac{h_0(t) \exp\{\sum_{i=1}^{p} \beta_i \mathbf{X_i^1}\}}{h_0(t) \exp\{\sum_{i=1}^{p} \beta_i \mathbf{X_i^2}\}} = \exp\{\sum_{i=1}^{p} \beta_i (\mathbf{X_i^1} - \mathbf{X_i^2})\}$$

Su interpretación es similar a la "*razón de ventajas*" *(odds ratio)* de una regresión logística, es decir, si la razón de riesgo es cercana a uno significa que no existen diferencias entre un diagnóstico u otro, pero si es significativamente diferente de uno se puede concluir que un valor del diagnóstico aumenta el riesgo de presentar un evento en un determinado tiempo. Estos cambios los especifican los parámetros asociados a cada factor introducido al modelo.

Después de ajustar el modelo de regresión de Cox, es importante verificar si se ajusta adecuadamente a los datos, para lo que se utilizan distintas clases de residuos que sirven para verificar las características propias de dicho modelo. Como última fase del análisis, es muy importante evaluar también el impacto de cada observación en el ajuste del modelo ya que puede darse el caso en que estas produzcan un fuerte impacto en todo el vector de parámetros o en particular en cada una de las covariables que componen el modelo. En cualquiera de estos dos casos es importante, después de localizar estas observaciones, tratar de determinar a qué se debe su gran impacto en el modelo.

Técnicas parámetricas que incluyen covariables

Finalmente, existen procedimientos que toman en consideración ajustar un modelo con una distribución conocida, estos son llamados *modelos paramétricos*. Normalmente se consideran distribuciones como la exponencial, la Weibull o la Gompertz [26] y al igual que el modelo de Cox, estos requieren del cumplimiento del supuesto de riesgos proporcionales. Veamos algunas características de estos modelos parámetricos.

- **Modelo parámetrico exponencial**

  En este caso tendremos que la función de densidad para $\theta > 0$ y $t \geq 0$ viene dada por:

$$f(t) = \theta \mathsf{e}^{-\theta t}$$

  Tendremos entonces las siguientes expresiones para las funciones de sobrevivencia, riesgo y riesgo acumulado:

$$S(t) = \int_t^\infty \theta \mathsf{e}^{-\theta u} \mathtt{du} = -\mathsf{e}^{-\theta u}\Big|_t^\infty = \mathsf{e}^{-\theta t} \quad , \quad h(t) = \frac{f(t)}{S(t)} = \frac{\theta \mathsf{e}^{-\theta t}}{\mathsf{e}^{-\theta t}} = \theta$$

$$\text{y} \quad \Lambda(t) = \int_0^t h(u) \mathtt{du} = \int_0^t \theta \mathtt{du} = \theta t.$$

  El hecho de que la función de riesgo sea constante, implica que el riesgo de presentar la falla no cambia a lo largo del tiempo. Esta característica de la función de riesgo de la exponencial está relacionada con la propiedad de pérdida de memoria de esta distribución, ya que la probabilidad de fallar en un intervalo $[t,\ t + \delta t]$, no depende de la sobrevivencia previa a este intervalo. Esta función es de uso común en contextos de confiabilidad.

- **Modelo parámetrico de Weibull**

  Una de las parametrizaciones de la función de densidad de la distribución Weibull para $\lambda > 0$, $\gamma > 0$ y $t \geq 0$ es la siguiente:

$$f(t) = \gamma \lambda t^{\gamma - 1} \mathsf{e}^{-\lambda t^\gamma}.$$

Tendremos las siguientes expresiones para las funciones de sobrevivencia, riesgo y riesgo acumulado:

$$S(t) = \int_t^\infty \gamma\lambda u^{\gamma-1}\mathtt{e}^{-\lambda u^\gamma}\mathtt{du} = -\mathtt{e}^{-\lambda u^\gamma}\Big|_t^\infty = \mathtt{e}^{-\lambda t^\gamma} \qquad h(t) = \frac{f(t)}{S(t)} = \gamma\lambda u^{\gamma-1}$$

$$\Lambda(t) = \int_0^t h(u)\mathtt{du} = \int_0^t \gamma\lambda u^{\gamma-1} = \lambda t^\gamma.$$

Notemos que:

$$h'(t) = (\gamma-1)\gamma\lambda t(\gamma-2) \text{ por lo que tenemos que } h'(t)\begin{cases} > 0, & \text{si } \gamma > 1 \\ < 0, & \text{si } \gamma < 1 \\ = 0, & \text{si } \gamma = 1. \end{cases}$$

Por lo que concluimos que para $\gamma > 1$ este modela riesgos monótonos crecientes, si $\gamma < 1$ modela riesgos monótonos decrecientes y si $\gamma = 1,\ h(t) = \lambda$ y la distribución Weibull se transforma en la exponencial.

Este es, probablemente, el modelo paramétrico más usado para datos de sobrevivencia, siendo ampliamente utilizado en aplicaciones biomédicas.

- **Modelo parámetrico de Gompertz**

La función de densidad de esta distribución para $\psi > 0$, $\lambda > 0$ y $t \geq 0$ viene dada por:

$$f(t) = \lambda\psi^t \exp\left\{-\frac{\lambda}{\log(\psi)}(\psi^t - 1)\right\}$$

Tendremos entonces las siguientes expresiones para las funciones de sobrevivencia, riesgo y riesgo acumulado:

$$S(t) = \exp\left\{-\frac{\lambda}{\log(\psi)}(\psi^t - 1)\right\} \quad , \quad h(t) = \frac{f(t)}{S(t)} = \lambda\psi^t$$

$$\text{y} \quad \Lambda(t) = \int_0^t h(u)\mathtt{du} = \frac{\lambda(\psi^t - 1)}{\log(\psi)}.$$

Notemos que si:

$$\psi > 1 \Rightarrow \lim_{t\to\infty} h(t) = \infty \text{ y } h(0) = \lambda \ \text{ y } \ \psi < 1 \Rightarrow \lim_{t\to\infty} h(t) = 0.$$

Entonces tendremos que si $\psi > 1$ se modelan riegos que son distintos de cero, toman el valor λ al inicio del estudio y son infinitos cuando t es grande. Para $\psi < 1$ estaremos modelando riesgos que inician en λ y eventualmente desaparecen.

Los resultados obtenidos por estos modelos son interpretables de forma similar a los producidos por el modelo de Cox. A pesar de que no son tan usados como el modelo semiparamétrico de riesgos proporcionales, se ha podido comprobar que si la función de sobrevivencia se ajusta adecuadamente a la distribución elegida, estos brindan resultados más exactos y con menores errores estándar que el modelo de Cox.

Capítulo 2

Eventos Recurrentes

Hasta este momento hemos asumido que el evento de interés puede ocurrir solo una vez en el transcurso de la investigación; sin embargo, existen muchos escenarios en los que el sujeto puede experimentar un evento varias veces sobre el seguimiento. Los procesos que generan estos eventos repetidos en el tiempo se denominan *procesos de sucesos recurrentes* y los datos que proporcionan son llamados *eventos recurrentes*.

Algunos ejemplos de eventos recurrentes son:

- Reaparición de ciertos tipos de tumores.
- Cantidad de reclamaciones de garantía sobre la compra de automóviles.
- Cantidad de veces que un estudiante repite una materia.

Una característica definitoria de este tipo de eventos es que las ocurrencias no pueden producirse de forma simultánea en un mismo sujeto lo que implica que existe una ordenación lógica de las mismas. Esta característica los diferencia de la ocurrencia de múltiples eventos, los cuales sí pueden coincidir en el tiempo sobre un mismo individuo. Otra característica de estos es que no suelen utilizarse datos que presenten censura por la izquierda, ya que podríamos estar considerando entonces observaciones donde ya ocurrió el evento y no fue observado y tales observaciones no nos aportarían información sobre la ocurrencia del mismo.

En este capítulo veremos a grandes rasgos en qué consiste el análisis de eventos recurrentes y algunos métodos empleados en dicho análisis.

2.1. Análisis de Eventos Recurrrentes

A partir de 1980 comenzó el estudio en el campo del análisis de datos de eventos recurrentes y durante las últimas décadas se han producido muchos avances estadísticos en él, llegándose a proponer varios enfoques para el análisis de este tipo de datos.

Los objetivos frecuentes en estos análisis implican, en primera medida, comprender y describir los procesos de eventos individuales, identificar y caracterizar la variación a través de una

muestra de procesos, comparar grupos de procesos y por último, determinar la relación entre los predictores relevantes y la tasa en la que los eventos están ocurriendo.

Desde un punto de vista estadístico el análisis de eventos recurrentes presenta dos desafíos importantes, estos son la correlación intraindividual y las covariables que varían en el tiempo.

Es conocido como *correlación intraindividual* el fenómeno de que los eventos recurrentes en un sujeto estén relacionados. Esta tiene dos fuentes posibles de aparición: la dependencia de ocurrencia y la heterogeneidad individual. La dependencia de ocurrencia se refiere a la posibilidad que la ocurrencia de un evento modifique la tasa de eventos posteriores en el individuo; por ejemplo, a medida que le ocurre el primer ataque cardíaco a un sujeto, las posibilidades de que ocurra un segundo ataque cardíaco aumentan porque durante el primero se daña una parte del corazón. Por su parte, la correlación debida a la heterogeneidad individual se refiere a la situación en la que algunos sujetos son más propensos a experimentar un mayor (o menor) número de eventos que otros sujetos debido a razones desconocidas, no medidas o inconmensurables. Un ejemplo de esta última puede ser un estudio que mide el número de crisis respiratorias en un grupo de individuos y en el cual no se pregunta por el consumo de tabaco. Es probable que los fumadores tengan un patrón diferente al de los no fumadores, lo que daría como resultado una heterogeneidad entre los sujetos que no puede atribuirse a ningún factor conocido, ya que no se registró el estado del tabaquismo.

Se debe tener en cuenta que el ajuste adecuado de la correlación intraindividual, cualquiera sea la fuente de aparición, es esencial para corregir la estimación del error estándar. Lo anterior se debe a que si tratamos la observación correlacionada como no correlacionada exageraríamos la cantidad de información proporcionada por cada observación, lo que nos conduciría a estimar de forma incorrecta dichos errores.

El otro desafío relacionado con el análisis de eventos recurrentes es cómo lidiar con las covariables que varían en el tiempo. En muchos estudios hay algunas covariables que están sujetas a cambios con el tiempo. Un ejemplo de lo anterior en el caso del manejo del asma sería notar que la dosis y el tipo de medicamentos que se le administran al individuo están sujetos a cambios durante el transcurso del tiempo, lo que tiene un efecto directo sobre el resultado.

2.1.1. Deficiencias de los métodos tradicionales para el Análisis de Eventos Recurrentes

Actualmente, y a pesar de que existen varias técnicas poderosas disponibles para el análisis de datos de eventos recurrentes, la mayoría de los investigadores todavía están usando técnicas estadísticas tradicionales, como la prueba t, la regresión logística y la regresión lineal múltiple, para analizar sus preguntas de investigación donde el resultado de interés es de naturaleza recurrente. Analicemos a continuación algunas dificultades al actuar de esta forma.

Comencemos analizando la *prueba t (t-test)* la cual se usa generalmente para comparar el número de eventos recurrentes entre dos poblaciones en estudios de cohortes. Generalmente en

los estudios se observa que existen sujetos que son más propensos a experimentar un mayor número de eventos que otros, lo que puede distorsionar la suposición de normalidad y dar lugar a una estimación inadecuada del error estándar. En los casos anteriores podríamos optar por utilizar la contraparte no paramétrica mejorada de la prueba t, conocida como la *prueba Wilcoxon de suma de rangos (Wilcoxon's Ranksum)* la cual no requiere el supuesto de normalidad.

Otra dificultad al utilizar la prueba t es que como ninguna de las observaciones del estudio son exactas se tiene como única forma de evaluar esta "confusión" hacer un análisis de subgrupos, lo cual solo sería factible hasta dos o tres factores de "confusión". Si tenemos un mayor número de estos factores, la solución sería realizar una regresión lineal múltiple, pero en el caso de análisis de eventos recurrentes, el supuesto de que los residuos siguen una distribución normal no se satisface. Además, esta regresión asume un riesgo uniforme en todos los eventos, lo que sabemos que en el caso que nos ocupa no se cumple necesariamente ya que el riesgo de eventos posteriores puede ser diferente del riesgo de eventos anteriores.

Por último, recordemos que esta prueba no nos ofrece forma de lidiar con covariables que dependen del tiempo, lo que consiste en otra dificultad con este procedimiento.

Otro método tradicional utilizado, incluso con mayor frecuencia que el anterior, es la regresión logística. En esta se divide a todos los sujetos en dos grupos como aquellos que experimentaron cualquier evento y aquellos que no experimentaron ningún evento y entonces la proporción de sujetos con y sin eventos se compara con grupos de tratamiento y control ajustando por las variables de confusión. Este tratamiento analítico de los datos hace que un sujeto que experimentó solo un evento durante el seguimiento sea igual a otro sujeto que experimentó más de un evento. En otras palabras, no se distingue a los sujetos con diferentes número de eventos y los agrupa a todos en el mismo grupo ignorando el número de eventos en el análisis. Esta consiste en una descripción extremadamente inadecuada de los procesos de eventos recurrentes y resulta la más crítica deficiencia de esta técnica para analizar este tipo de datos.

Como otra deficiencia podemos señalar también que la regresión logística no puede acomodar las covariables dependientes del tiempo de forma natural en el análisis lo que conduce a resultados incompletos y/o inapropiados en la descripción del proceso.

La regresión de riesgo proporcional de Cox es otra técnica muy utilizada, específicamente cuando tenemos que analizar datos de eventos recurrentes y estos traen consigo información disponible sobre el tiempo en que ocurrió el evento. Como vimos anteriormente, es una técnica de análisis de sobrevivencia lo que hace que funcione mejor que la regresión logística en los casos en que tengamos la información del tiempo y que dicha información juegue un papel importante en las preguntas de investigación. A pesar de que tiene la capacidad de manejar variables que varían en el tiempo no es apropiada para el análisis de eventos recurrentes ya que esta usa la información hasta el primer evento solamente, quedando inutilizada toda la información después de este. Debido a lo anterior, la utilización de esta técnica puede conducir a una evaluación inexacta de la eficacia de un tratamiento. En particular, esta puede subestimar sustancialmente los beneficios potenciales en términos del evento prevenido por un tratamiento.

Hasta aquí hemos mencionado algunos métodos tradicionales que se utilizan para analizar datos de eventos recurrentes de forma inapropiada. Existen diferentes estudios donde se demuestra la ineficacia de estos métodos tradicionales en los casos que nos competen. Lo anterior lo hacen exponiendo la comparación de los resultados obtenidos al analizar un conjunto de datos con estas características empleando los métodos anteriormente citados y métodos adaptados específicamente para estos datos. Como era de esperar mediante los métodos tradicionales se obtienen resultados, en el mejor de los casos, inexactos, si no es que totalmente erróneos.

Un ejemplo de lo anterior es el estudio realizado por RJ Glynn et al. [9] donde analizan unos datos de ensayos clínicos de tres formas diferentes con el objetivo de evaluar el efecto de una sustancia en una bacteria. En este estudio aleatorizan a 153 pacientes en dos grupos, uno de tratamiento y otro de control, posteriormente realizan el experimento, recolectan la información y realizan los análisis de los datos. Al emplear métodos tradicionales no obtienen ninguna diferencia significativa entre los dos grupos mientras que al emplear métodos para eventos recurrentes se observó una diferencia sustancial entre los dos grupos.

A pesar de lo anterior, al hacer revisiones de la literatura se encuentra que son muy pocos los autores que no utilizan métodos tradicionales al realizar investigaciones que tienen como variables de interés un evento recurrente. Las posibles explicaciones de por qué sucede esto a pesar de la disponibilidad de alternativas apropiadas son: o bien no conocen estas técnicas ya que la mayoría de estas se discuten en literatura específica, generalmente difíciles de entender para aquellos ajenos al área o es posible que no se disponga de directrices claras con respecto a la selección de una alternativa adecuada, basada en la pregunta de investigación y la naturaleza de los datos.

2.2. Métodos de Análisis de Eventos Recurrentes

Revisadas anteriormente las dificultades que trae consigo analizar con técnicas inapropiadas estos datos, así como los retos que aún se presentan en este tema, dedicamos esta sección a describir varios métodos de análisis de eventos recurrentes. Para facilitar la comprensión de los mismos y sus características los clasificaremos en dos categorías dependiendo de si se basan o no en modelos lineales generalizados.

2.2.1. Métodos basados en Modelos Lineales Generalizados

Las **variables de recuento** son aquellas que determinan el número de sucesos o eventos que ocurren en una misma unidad de observación en un intervalo espacial o temporal definido. Se caracterizan principalmente por su naturaleza discreta y sus valores no negativos.

Ejemplos de variables de recuento pueden ser:

- Número de patentes solicitadas por una empresa en 12 meses.
- Número de artículos publicados por una revista.

Los modelos para variables de este tipo son conocidos como *modelos de datos de recuento (o de conteo)* y son un caso particular de modelización lineal. Estos nos permiten considerar y analizar el comportamiento de las variables de conteo frente a los valores del conjunto de variables explicativas. Dichos modelos se pueden utilizar para analizar eventos recurrentes en situaciones donde la información sobre el tiempo en que tuvieron lugar los eventos de interés no está disponible, o simplemente, no resulta relevante para la investigación.

Entre los diferentes enfoques que existen para tratar estas situaciones, los dos métodos más comúnmente usados son, la Regresión de Poisson y la Regresión Binomial Negativa bajo ciertas restricciones en el espacio parametral [31], ambos miembros de la familia de modelos lineales generalizados.

Regresión de Poisson

Sea Y una variable aleatoria con distribución de Poisson, entonces su función de probabilidad viene dada por:

$$P(Y = y) = \frac{e^{-\mu}\ \mu^y}{y!} \qquad \mu > 0 \qquad y = 0, 1, 2, ...$$

y su media y varianza vienen dadas por:

$$\mathbb{E}(Y) = \mu \qquad\qquad \texttt{Var}(Y) = \mu$$

Para construir el modelo donde la variable independiente sigue una distribución Poisson, especificamos el parámetro μ_i como una forma funcional de las variables explicativas $X_1, ..., X_p$. Haciendo uso de la función de enlace canónica para la formación del modelo, o sea, $\mu_i = \exp(x_i\ \beta)$ obtendríamos que:

$$E(Y_i \mid x_i) = \exp(\beta_0 + \beta_1 x_{i1} + \ ... \ + \beta_p x_{ip})$$

Por tanto la distribución de Poisson condicionada a las variables explicativas $X_{i's}$ viene dada por:

$$P(Y_i = y_i \mid x_i) = \frac{e^{-\mu_i(x_i)}\mu_i(x_i)^{y_i}}{y_i!} \qquad y_i = 0, 1, 2, ..., p$$

donde

$$E(Y_i \mid x_i) = \mu_i(x_i) = \mu_i(x_{i1}, ..., x_{ip}) = \exp(\beta_0 + \beta_1 x_{i1} + \ ... \ + \beta_p x_{ip})$$

La formulación anterior se conoce como *Modelo de Regresión de Poisson*, el cual modela el número de ocurrencias de un evento o la tasa de eventos en función de algunas variables explicativas.

Si la especificación para la distribución condicional de la variable de respuesta, así como la de la media condicional, es correcta, y bajo el supuesto de que se tienen observaciones independientes, entonces se puede utilizar la siguiente función de logverosimilitud $L(.)$ para mediante el método de máxima verosimilitud obtener estimadores consistentes β:

$$L(\beta) = \log\left\{\prod_{i=1}^{p} \frac{e^{-\mu_i(x_i)}\mu_i(x_i)^{y_i}}{y_i!}\right\} = \log\left\{\frac{e^{-\sum_{i=1}^{p}\mu_i(x_i)}\prod_{i=1}^{p}\mu_i(x_i)^{y_i}}{\prod_{i=1}^{p} y_i!}\right\}$$

$$L(\beta) = (...) = \sum_{i=1}^{p}\{-\mu_i(x_i) + y_i \log\{\mu_i(x_i)\} - \log(y_i!)\} = \sum_{i=1}^{p}\{y_i x_i'\beta - \exp\{x_i'\beta\} - \log(y_i!)\}$$

Se ha demostrado que, aunque el supuesto de la igualdad entre la media y la varianza condicionales para este modelo no se cumpla, situación que es muy común, el estimador puntual de β es aún válido, pero no así el estimador de su error estándar, y por tanto las inferencias respecto de β tampoco lo sería [11]. Para los casos en que esto suceda se han propuesto algunas alternativas que suavizan este supuesto pero mantienen el supuesto de una distribución condicional Poisson. En particular se ha propuesto el uso de errores estándar robustos, algo también conocido como método de estimación de pseudomáxima verosimilitud [10], el empleo de un enfoque de cuasiverosimilitud [19] o la utilización de errores estándar bootstrap [8].

Un problema común en la regresión de Poisson es, por ejemplo, que los datos presenten una frecuencia de ceros no consistente con el modelo. Este problema se vuelve relevante cuando el exceso de ceros es causado porque en realidad hay dos procesos subyacentes:

1- Un fenómeno que determina la presencia de un cero frente a un valor positivo.

2- Una vez que se alcanza un valor positivo, hay otro fenómeno que determina el conteo de eventos que se producen.

Un ejemplo clásico sería la distribución de cigarrillos fumados en una hora por miembros de un grupo en el que algunos individuos no son fumadores. En estos casos, existen alternativas para modelar este problema como la regresión cero-inflada o la regresión binomial negativa cero-inflada [16].

Otra situación a la que frecuentemente nos enfrentamos en los modelos para datos de recuento es que no se cumpla la igualdad entre la media y la varianza, siendo esta última mayor, fenómeno conocido como **sobredispersión**. En este caso, no se validarían las hipótesis de la regresión de Poisson, lo que ocasionaría que se subestime el error estándar dando lugar a p-valores sesgados e intervalos de confianza muy estrechos. Una razón común para que suceda esto es la omisión de variables explicativas o de observaciones relevantes, por lo que lo primero que debe hacerse es revisar los datos y las variables consideradas en el análisis. Sin embargo, el problema de sobredispersión puede persistir y, en algunas circunstancias, puede resolverse usando otros modelos como el de regresión binomial negativa o un modelo de cuasiverosimilitud.

Como hemos visto, a pesar de que el modelo de referencia en estudios de variables de recuento es el modelo de regresión de Poisson este presenta varios problemas a la hora de tratar con situaciones que comúnmente encontramos en datos reales. Una forma de tratar lo

anterior sería especificar una distribución que permita un modelado más flexible, donde una de las opciones es el modelo de regresión binomial negativa, el cual analizaremos a continuación.

Regresión Binomial Negativa

Partiendo del modelo de regresión de Poisson podemos obtener una generalización del *modelo de regresión Binomial Negativa* mediante la incorporación de un término de perturbación, una aleatoriedad en el parámetro μ_i [5]; o sea

$$\mu_i^* = \exp(x_i\beta + \epsilon_i) = \mu_i \ \exp(\epsilon_i)$$

donde el término de perturbación ϵ_i sigue una distribución Gamma.

Su función de probabilidad será:

$$P(Y = y_i|x_i) = \frac{\Gamma(y_i + v_i)}{\Gamma(y_i + 1)\ \Gamma(v_i)} \left(\frac{v_i}{v_i + \mu_i}\right)^{v_i} \left(\frac{\mu_i}{\mu_i + v_i}\right)^{y_i}$$

con $\mu_i = \mathbb{E}[Y_i|x_i] = \exp(x_i\ \beta)$ y definimos $v_i = (1/\alpha){\mu_i}^t$ donde $t = 0, 1$.

La especificación final del modelo depende del valor que tome t, o sea:

- Si $t = 0$, $v_i = (1/\alpha)$ tendríamos:

$$\mathbb{E}(Y_i|x_i) = \exp(x_i\beta) \qquad \texttt{Var}(Y_i|x_i) = (1 + \alpha)\exp(x_i\beta)$$

- Si $t = 1$, $v_i = (1/\alpha)\mu$ tendríamos:

$$\mathbb{E}(Y_i|x_i) = \exp(x_i\beta) \qquad \texttt{Var}(Y_i|x_i) = \exp(x_i\beta)(1 + \alpha\ \exp(x_i\beta))$$

En estos modelos si $\alpha > 0$ entonces $\texttt{Var}(Y_i \mid x_i) \geq \mathbb{E}(Y_i \mid x_i)$, lo que sugiere que los datos presentan sobredispersión.

Para el caso en que $t = 0$ (*BN-I*) se tiene una relación lineal entre la media y la varianza, mientras que cuando $t = 1$ (*BN-II*) esta relación será cuadrática.

La funciones de logverosimilitud para ambos casos después de algunas manipulaciones algebraicas de la función de densidad, son las siguientes:

BN-I:

$$L(\beta, \alpha) = \sum_{i=1}^{p} \left[-\log(y_i!) + \sum_{j=1}^{y_i} \log(\alpha y_i + \mu_i - \alpha_j) - \left(\frac{\mu_i}{\alpha} + y_i\right)\log(1 + \alpha) \right]$$

BN-II:

$$L(\beta, \alpha) = \sum_{i=1}^{p} \left[-\log(y_i!) + \sum_{j=1}^{y_i} \log(\alpha y_i + 1 - \alpha_j) + y_i \log(\mu_i) - \left(\frac{1}{\alpha} + y_i\right)\log(1 + \alpha\mu_i) \right]$$

Además de los modelos anteriores, algunos autores como [5] proponen un modelo binomial negativo más general, el denominado *Hipermodelo Binomial Negativo K* (BN K), en el cual se tiene que $\texttt{Var}(Y_i \mid x_i) = \mu_i + \alpha\ \mu_i^{2-k}$, donde k es un parámetro fijo con valor no negativo.

El método más usual para la estimación de los parámetros en este modelo, al igual que en el anterior, es el método de máxima verosimilitud, en el cual, debido a la complejidad de las ecuaciones de logverosimilitud obtenidas, se suele recurrir a procedimientos numéricos para la resolución de las mismas. La mayoría de los investigadores proponen el método de Newton-Raphson o el método Fisher *scoring*, dependiendo de como haya sido concebido el modelo. Ambos métodos son usados comúnmente para el análisis de datos tanto Poisson como Binomiales Negativos.

El uso de la regresión binomial negativa, como cualquier otro método estadístico, presenta ventajas y desventajas. Entre sus ventajas podemos citar que, cuando los datos no cumplen algunos de los supuestos, este da un mejor ajuste en comparación con la regresión de Poisson. También podemos mencionar que debido a que la varianza de la distribución Binomial Negativa es siempre mayor que la varianza de la distribución de Poisson, esta regresión permite más variabilidad que la regresión de Poisson. A pesar de que como hemos visto presenta cierta "superioridad" frente a la regresión de Poisson, también tiene algunas limitaciones. Entre las limitaciones tenemos que generalmente se utiliza la distribución Gamma para modelar el término de perturbación de los individuos, ya que es fácil de entender y viene implementada en los softwares estadísticos, no obstante esta no siempre es una distribución adecuada, por lo que es aconsejable probar más de una distribución para describir este término.

Modelos truncados en cero

Los modelos truncados implican que en algún punto del "recorrido" de la variable, un determinado valor está totalmente ausente. Específicamente, cuando estamos trabajando con eventos recurrentes es común al recolectar los datos tener un sesgo ya que casi nunca se consideran individuos con cero eventos. Lo anterior se debe a que como estamos estudiando la ocurrencia repetida de cierto evento no es de interés aquellos individuos que no lo presentaron. Bajo estas ideas se tienen los *modelos truncados en cero*.

Este tipo de modelos no admite conteos ceros, por lo que la distribución no debe tener este valor en su recorrido para poder modelar los datos adecuadamente. Pensando en las distribuciones anteriormente utilizadas, debemos modificarlas para llegar a sus versiones truncadas, pero debemos tener en cuenta no solo que no admitan ceros si no también que la suma de las probabilidades de los valores sea 1 para que sigan siendo distribuciones de probabilidad.

Veamos las funciones de distribución que se obtendrían en cada caso:

- **Poisson Cero Truncado**

$$f(Y_i = y_i \mid x_i, y_i > 0) = \frac{f(Y_i = y_i,\ y_i > 0 \mid x_i)}{f(y_i > 0 \mid x_i)} = \frac{e^{-\mu_i(x_i)}\mu_i(x_i)^{y_i}}{y_i!\ (1 - \exp\{-\mu_i(x_i)\})}$$

- **Binomial Negativa Cero Truncado**

$$f(Y_i = y_i \mid x_i, y_i > 0) = \frac{f(Y_i = y_i,\ y_i > 0 \mid x_i)}{f(y_i > 0 \mid x_i)} = \frac{\dfrac{\Gamma(y_i + v_i)}{\Gamma(y_i + 1)\ \Gamma(v_i)} \left(\dfrac{v_i}{v_i + \mu_i}\right)^{v_i} \left(\dfrac{\mu_i}{\mu_i + v_i}\right)^{y_i}}{1 - \left(\dfrac{v_i}{v_i - \mu_i}\right)^{v_i}}$$

Ajuste y selección de los modelos

Una vez elegido el modelo de regresión es necesario evaluar si el mismo tiene un buen ajuste y si es el indicado para estudiar nuestros datos. Lo anterior implica analizar residuos, estadísticos de bondad de ajuste y realizar pruebas para corroborar la selección del modelo. Veamos a continuación aspectos generales de estos procedimientos.

- **Estadísticas de bondad de ajuste**

  Los modelos basados en modelos lineales generalizados para medir la bondad del ajuste suelen emplear estadísticos que conocemos como la devianza, el estadístico de Chi Cuadrado de Pearson y el coeficiente de determinación R^2. Veamos las expresiones de estos.

  - Función de devianza: Se define de forma general como:

$$D(y; \widehat{\mu}) = 2\{L(y; y) - L(\widehat{\mu}; y)\}$$

    donde $L(y; y)$ es la logverosimilitud del modelo saturado en la cual se tienen n parámetros, o sea, uno por cada observación y $L(\widehat{\mu}; y)$ es la logverosimilitud del modelo a ser estimado.

    Si el modelo es correcto y está bien definido el estadístico anterior se distribuye asintóticamente según una $\chi^2_{(n-p)}$ donde $n - p$ son los grados de libertad. Para probar entonces la adecuación de este modelo debemos comparar el valor del estadístico con el percentil adecuado de la distribución obtenida. Si el p-valor resultante es menor que el nivel de significación establecido, se rechaza la hipótesis nula la cual en este caso es $H_0 : D = 0$.

    Resumiendo, lo podemos interpretar como que un valor pequeño de esta función indica que para el número de parámetros que se están considerando se obtiene un ajuste tan bueno como cuando se ajusta con el modelo saturado.

    Para los casos vistos anteriormente tendríamos:

    - Regresión de Poisson:

$$D(y; \widehat{\mu}) = 2 \sum_{i=1}^{n} \{y_i \log(y_i / \widehat{\mu_i}) - (y_i - \widehat{\mu_i})\}$$

- Regresión Binomial Negativa:

$$D(y;\widehat{\mu}) = 2\sum_{i=1}^{n}\{y_i \log(y_i/\widehat{\mu_i}) - (y_i + 1/\widehat{\alpha})/\widehat{\alpha}\ \log((1+\widehat{\alpha y_i})/(1+\widehat{\alpha}\widehat{\mu_i}))\}$$

- Coeficiente de determinación R^2: Este se define como la reducción proporcional en la incertidumbre a medida que se incluyen las variables explicativas, y bajo ciertas condiciones, se puede interpretar como la varianza explicada por el modelo ajustado. Este coeficiente toma valores en el intervalo $(0, 1)$ y viene dado por:

$$R^2 = 1 - \frac{D(y;\widehat{\mu})}{D(y;\widehat{\mu_0})}$$

donde $D(y;\widehat{\mu})$ y $D(y;\widehat{\mu_0})$ son las funciones de devianza correspondientes al modelo ajustado y al modelo nulo respectivamente. Entiéndase por modelo nulo aquel que no incluye ninguna variable explicativa.

- Estadístico Chi-cuadrado de Pearson: Se define como:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{\mathtt{V}(\widehat{\mu_i})}$$

donde $\mathtt{V}(\widehat{\mu})$ es la función varianza estimada para la distribución de la variable respuesta.

Para los casos particulares vistos tendríamos:

- Regresión de Poisson:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{\mathtt{V}(\widehat{\mu_i})} = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{\widehat{\mu_i}}$$

- Regresión Binomial Negativa:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{\mathtt{V}(\widehat{\mu_i})} = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{(\widehat{\mu_i} + \widehat{\alpha}\widehat{\mu_i})^2}$$

- **Análisis de residuos**

Los residuos expresan la discrepancia entre una observación y su valor ajustado y, al igual que en los modelos de regresión lineal, son utilizados para verificar la adecuación del modelo. Estos pueden indicar la presencia de valores anómalos o discordantes que puedan requerir una investigación más detallada. Los residuos más comúnmente usados son los siguientes:

- Residuos básicos: Se definen como la diferencia entre el valor observado de la variable respuesta y su valor ajustado, o sea:

$$r_i^{\mathbf{B}} = y_i - \widehat{y}_i \qquad \text{para } i = 1, ..., n$$

- Residuos de Pearson: Estos representan la contribución individual de cada observación al estadístico χ^2 de Pearson, o sea:

$$X^2 = \sum_{i=1}^{n} (r_i^{\mathrm{P}})^2$$

Para los casos particulares que hemos tratado estos residuos se definen como:

- Regresión de Poisson:

$$r_i^{\mathbf{P}} = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{\mu_i}}} \qquad \text{para } i = 1, ..., n$$

- Regresión Binomial Negativa:

$$r_i^{\mathbf{P}} = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{\mu_i} + \widehat{\alpha}\widehat{\mu_i}^2}} \qquad \text{para } i = 1, ..., n$$

- Residuo de Pearson estudentizado: Estos captan mejor la variabilidad de los datos ya que tienen en cuenta al valor de h_i, que representa al i-ésimo elemento de la matriz proyección y el cual es útil para medir la influencia de la i-ésima observación. Para los casos tratados tendremos:

- Regresión de Poisson:

$$r_i^{\mathbf{PS}} = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{\mu_i}(1 - h_i)}} \qquad \text{para } i = 1, ..., n$$

- Regresión Binomial Negativa:

$$r_i^{\mathbf{PS}} = \frac{y_i - \widehat{\mu_i}}{\sqrt{(1 - h_i)(\widehat{\mu_i} + \widehat{\alpha}\widehat{\mu_i}^2)}} \qquad \text{para } i = 1, ..., n$$

- Residuos de la devianza: Estos son interpretados como una medida de discrepancia en los ajustes de un modelo lineal generalizado ya que cada observación contribuye en el valor de la devianza, de manera que la suma de contribuciones de todas las observaciones toma por valor la devianza, o sea estos cumplen que:

$$D(y; \widehat{\mu}) = \sum_{i=1}^{n} \left(r_i^{\mathrm{D}}\right)^2$$

Para los casos particulares que hemos visto, estos tienen las siguientes expresiones:

- Regresión de Poisson:

$$r_i^{\mathbf{D}} = \mathrm{sg}(y_i - \widehat{\mu_i}) \left\{ 2 \left[y_i \log\left(\frac{y_i}{\widehat{\mu_i}}\right) - (y_i - \widehat{\mu_i}) \right] \right\}^{1/2} \qquad \text{para } i = 1, ..., n$$

- Regresión Binomial Negativa:

$$r_i^{\mathbf{D}} = \mathrm{sg}(y_i - \widehat{\mu_i}) \left\{ 2 \left[y_i \log\left(\frac{y_i}{\widehat{\mu_i}}\right) - (y_i + \widehat{\alpha}) \log\left(\frac{y_i + \widehat{\alpha}}{\widehat{\mu_i} + \widehat{\alpha}}\right) \right] \right\}^{1/2} \qquad \text{para } i = 1, ..., n$$

- Residuos de Anscombe: Estos residuos buscan normalizar el residuo básico con el objetivo de que la heterogeneidad y los valores atípicos se puedan identificar rápidamente. Se definen basados en una función $A(y)$ en lugar de y y se trata de garantizar que la distribución de $A(y)$ se asemeje tanto a la distribución normal como sea posible.

  Sus expresiones para los casos que hemos visto son:

  - Regresión de Poisson:

$$r_i^{\mathbf{A}} = \frac{\frac{3}{2}\left(y_i^{2/3} - \widehat{\mu_i}^{2/3}\right)}{\widehat{\mu_i}^{1/6}} \qquad \text{para } i = 1, ..., n$$

  - Regresión Binomial Negativa:

$$r_i^{\mathbf{A}} = \frac{\frac{3}{\widehat{\alpha}}\left\{(1 + \widehat{\alpha}\ y_i)^{2/3} - (1 + \widehat{\alpha}\ \widehat{\mu_i})^{2/3}\right\} + 3\left(y_i^{2/3} - \widehat{\mu_i}^{2/3}\right)}{2\left(\widehat{\mu_i} + \widehat{\alpha}\ \widehat{\mu_i}^2\right)^{1/6}} \qquad \text{para } i = 1, ..., n$$

- **Pruebas para la selección de modelos**

  Se utilizan pruebas comparativas como criterios para seleccionar el modelo más apropiado a los datos. Las principales pruebas empleadas son el Criterio de la Información de Akaike (AIC) y Criterio de Información Bayesiano (BIC), siendo los modelos que presentan valores menores los que indican un mejor ajuste. Estos criterios consisten en una serie de parametrizaciones alternativas, cada una de las cuales tiene como objetivo determinar un método para evaluar mejor el ajuste del modelo.

  - AIC: Este estadístico se define de la siguiente forma:

$$AIC = p - 2\log(\widehat{L})$$

    donde $\widehat{L}$ representa el valor máximo de la función de verosimilitud del modelo estimado y p el número de variables explicativas del mismo.

    En este se penaliza la cantidad de predictores con el término p, dado que al aumentar la cantidad de los mismos aumenta el valor de este estadístico y por tanto el modelo va perdiendo idoneidad. Lo anterior está en concordancia con el principio de parsimonia, el cual estipula que en igualdad de condiciones el modelo más sencillo, suele ser el mejor.

  - BIC: Este estadístico tiene la forma:

$$BIC = -2\log(\widehat{L}) + p\log(n)$$

donde $\widehat{L}$ y p representan lo mismo que en el caso anterior y n se refiere al número de observaciones.

Como particularidad tenemos que este da un mayor peso al término de ajuste $p\ \log(n)$ que el anterior.

Estimación e interpretación de los parámetros del modelo

En estos modelos se utilizan para la estimación de los parámetros generalmente el método de mínimos cuadrados o el método de máxima verosimilitud, siendo este el más adecuado ya que tiene las propiedades de consistencia y eficiencia asintótica.

Se define como estimador máximo verosímil de un parámetro a aquel valor que maximiza la probabilidad de observar una determinada muestra. Para hallar estos estimadores se calculan los máximos locales de la función de logverosimilitud. El procedimiento para lo anterior sería derivar la logverosimilitud respecto a los parámetros, β en nuestro caso, e igualar a cero las expresiones obtenidas. Posteriormente se despejan los valores de β y obtendríamos así las estimaciones de los parámetros.

En el procedimiento explicado anteriormnete y debido a la complejidad de las ecuaciones de logverosimilitud obtenidas, generalmente se debe recurrir a métodos numéricos, como los de Newton-Raphson y de Fisher *scoring*, para la obtención de los valores máximos.

Para desarrollar estos métodos necesitaremos la función score, la matriz hessiana y las matrices de información de Fisher esperada y observada para cada modelo. Tomando la función de logverosimilitud del modelo como $L(\beta)$, podríamos definir los anteriores de la siguiente forma:

- Función score: Es la primera derivada de la logverosimilitud respecto a los parámetros, o sea:
$$s(\beta) = \frac{\partial L(\beta)}{\partial \beta} = \sum_i s_i(\beta)$$
- Matriz Hessiana: Es una matriz cuadrada de $n \times n$ definida como
$$(H_L)_{i,j} = \frac{\partial^2 L}{\partial \beta_i\, \partial \beta_j}$$
- Matriz de información de Fisher observada:
$$\mathcal{I}_{obs}(\beta) = -\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^t}$$
- Matriz de información de Fisher esperada:
$$\mathcal{I}_{esp} = \ \text{Cov}\ (s(\beta)) = \sum_i \mathcal{I}_{esp_i}(\beta)$$

y se puede comprobar que:

$$\mathcal{I}_{esp}(\beta) = \mathbb{E}(\mathcal{I}_{obs}(\beta)) = \mathbb{E}\left[\frac{\partial^2 L(\beta)}{\partial\beta\partial\beta^t}\right]$$

Veamos dos ejemplos de las ecuaciones de iteración para estos métodos en los modelos estudiados.

- Newton-Raphson para el modelo de Regresión de Poisson

  En este caso, la estimación de los parámetros β para la i-ésima iteración vendrá dada por la ecuación siguiente:

$$\widehat{\beta}^{(k)} = \widehat{\beta^{(k-1)}} + (\mathtt{X}^k\ \mathtt{W}^{(k-1)}\ \mathtt{X})^{-1}\ \mathtt{X}^{(k)}(Y - \pi) \qquad k = 1, 2, ...$$

  donde la matriz hessiana será:

$$[H_{L(\beta)}]_{(i,j)} = \left[\frac{\partial^2 L(\beta)}{\partial\beta\partial\beta^k}\right]_{(i,j)} = -\mathtt{X}^{\mathtt{k}}\ \mathtt{W}\ \mathtt{X}$$

  con

$$\mathtt{W} = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_n \end{bmatrix} \qquad \text{y} \qquad \pi_i = \mathtt{e}^{\sum_{j=1}^{n+1}\beta_j x_{ij}} \qquad \text{para } i = 1, 2, ..., n$$

- Fisher *scoring* para el modelo de Regresión Binomial Negativo

  Este método es el que se utiliza con mayor frecuencia en los modelos lineales generalizados para estimar los parámetros.

  Sus iteraciones vienen definidas de la siguiente forma:

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + \mathcal{I}_{esp}(\widehat{\beta}^{(k)})\ s(\widehat{\beta}^{(k)}) \qquad k = 0, 1, 2, ...$$

  Este permite una estimación consistente únicamente del parámetro β y el parámetro de dispersión se incluirá como una constante conocida.

Ambos métodos iterativos pueden utilizar el siguiente criterio de parada:

$$\frac{\|\widehat{\beta}^{(k+1)} - \widehat{\beta}^{(k)}\|}{\|\widehat{\beta}^{(k)}\|} < \epsilon \qquad \text{para } \epsilon > 0 \ \text{ fijo}$$

A modo de comparación podemos decir que, para el mismo modelo, este último algoritmo generalmente requiere más iteraciones para estimar los parámetros que el de Newton-Raphson, no obstante los cálculos que se deben realizar en este son más simples comparados con aquellos que se realizan en el primer algoritmo planteado.

Una vez se ha obtenido el modelo adecuado haciendo uso de todos los criterios antes mencionados el proceso de modelado se cierra con la interpretación del modelo. Teniendo en cuenta que la ecuación del modelo estará expresada en términos multiplicativos, la interpretación de los parámetros se realiza en términos del factor de cambio en el valor esperado para un incremento unitario de las variables explicativas. En otras palabras, tendríamos que para un cambio de δ unidades en la variable explicativa x_j manteniendo el resto de variables constantes, el recuento esperado (en nuestro caso particular el número de eventos recurrentes que tienen lugar) se incrementa en un factor de $\exp\{\beta_j\delta\}$. Lo anterior nos sería util para comparar eventos en diferentes grupos, por ejemplo entre mujeres y hombres, pero presenta una gran dificultad, y es que al igual que en la regresión logística, no se utiliza el tiempo en que ocurren los eventos por lo que estaremos perdiendo exactitud en nuestras estimaciones.

2.2.2. Métodos no basados en Modelos Lineales Generalizados

Varios estadísticos han contribuido al desarrollo de modelos refinados para analizar desde la teoría del análisis de sobrevivencia los datos de eventos repetidos. Estas técnicas desarrolladas comparten puntos en común, y todas se basan en realizar inferencias con el objetivo de ajustar el modelo de riesgo proporcional mediante la combinación de todos los eventos dentro de un proceso de maximización integrado.

Sean T_{iq} el tiempo real del evento para el $q-$ésimo tiempo de experimentación de cierto evento para el individuo i, donde $q = 1, ..., Q$; C_{iq} el tiempo censurado con respecto al $q-$ésimo evento para el mismo individuo, y $t_{iq} = \texttt{min}(T_{iq},\ C_{iq})$ el tiempo de observación. Entonces, la función de verosimilitud parcial de un modelo con eventos recurrentes se puede escribir como [18]

$$L_p(\beta) = \prod_{i=1}^{d}\prod_{q=1}^{Q} \frac{\exp[x_{iq}^t(t)\ \beta]}{\sum_{l\in R(t_i)}\exp[x_{lq}^t(t)\ \beta]}, \quad \text{donde} \quad x_{iq}^t(t) = x^t(t_{iq}).$$

Cada uno de los modelos para eventos recurrentes dependen de sus propios supuestos y de la dependencia intraindividuos que se esté considerando. En consecuencia, en cada modelo se especifican diferentes conjuntos $R(t_i)$ y se proponen enfoques estadísticos diferentes para ajustar la correlación intraindividuos. Como resultado los modelos que manejan eventos recurrentes a menudo producen resultados sustancialmente diferentes. Dada tal variabilidad, comprender las características, las fortalezas y las limitaciones de cada modelo es esencial para emplear correctamente estas técnicas con el fin de analizar los datos de supervivencia con eventos recurrentes de una forma eficiente.

Antes de pasar a analizar algunos de los modelos de análisis de sobrevivencia para eventos recurrentes debemos mencionar que entre ellos existe una distinción importante que debe ser tomada en cuenta. Esta distinción agrupará los métodos en dos grupos, uno con los métodos para modelar las situaciones donde todos los eventos recurrentes en el mismo individuo se tratan como idénticos, y el otro para los casos donde algunos de los eventos recurrentes implican diferentes categorías de enfermedades y/o el orden en que ocurre la repetición de eventos se considera importante.

Modelo Estándar de Cox de Riesgos Proporcionales

Procesos de Conteo

El diseño general de datos para el enfoque del proceso de conteo para un conjunto de datos que involucra N sujetos se presenta en la Tabla 2.1.

i	**j**	$\mathbf{d_{ij}}$	$\mathbf{t_{ij0}}$	$\mathbf{t_{ij1}}$	$\mathbf{X_{ij1}}$	...	$\mathbf{X_{ijp}}$
1	1	d_{11}	t_{110}	t_{111}	X_{111}	...	X_{11p}
1	2	d_{12}	t_{120}	t_{121}	X_{121}	...	X_{12p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	r_1	d_{1r_1}	t_{1r_10}	t_{1r_11}	X_{1r_11}	...	X_{1r_1p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	1	d_{i1}	t_{i10}	t_{i11}	X_{i11}	...	X_{i1p}
i	2	d_{i2}	t_{i20}	t_{i21}	X_{i21}	...	X_{i2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	r_i	d_{ir_i}	t_{ir_i0}	t_{ir_i1}	X_{ir_i1}	...	X_{ir_ip}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	1	d_{N1}	t_{N10}	t_{N11}	X_{N11}	...	X_{N1p}
N	2	d_{N2}	t_{N20}	t_{N21}	X_{N21}	...	X_{N2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	r_N	d_{Nr_N}	t_{Nr_N0}	t_{Nr_N1}	X_{Nr_N1}	...	X_{Nr_Np}

Tabla 2.1: Diseño general de datos para el enfoque del proceso de conteo

El $i-$ésimo sujeto tiene r_i eventos recurrentes, d_{ij} denota el estado del evento (1 = falla, 0 = censurado). Para este sujeto en el $j-$ésimo intervalo de tiempo, t_{ij0} y t_{ij1} denotan las horas de inicio y finalización respectivamente y X_{ijk} denota el valor del $k-$ésimo predictor para el $i-$ésimo sujeto en el $j-$ésimo intervalo.

Una característica de estos datos es que los sujetos considerados no están restringidos a tener el mismo número de intervalos de tiempo, o sea, r_1 no tiene que ser igual a r_2 ni tampoco deben tener el mismo número de eventos recurrentes. También se caracterizan por que si el último intervalo de tiempo para un sujeto dado termina en censura, entonces el número de eventos recurrentes para este sujeto es $r_i - 1$ y los intervalos de tiempo anteriores, por lo

general, terminan con una falla.

Estos datos cumplen, además, que los tiempos de inicio y finalización pueden ser diferentes para diferentes sujetos y al igual que con cualquier dato de supervivencia, las covariables X_{ijk} pueden ser independientes o dependientes del tiempo para un sujeto determinado. Un ejemplo de esto último podrían ser si una de las $X's$ es "género" los valores que esta tome en los intervalos de tiempo observados para un sujeto determinado serán todos 1 o todos 0; en cambio, si analizamos otra variable, por ejemplo, una medida del nivel de estrés diario, es probable que los valores de esta variable para un mismo individuo varíen durante los intervalos de tiempo analizados.

El modelo que se utiliza normalmente para llevar a cabo el enfoque de los procesos de conteo es el modelo estándar de Cox de riesgos proporcionales, o sea:

$$h(t, \mathbb{X}) = h_0(t) \ \exp\left(\sum \beta_i X_i\right),$$

donde t corresponde al tiempo que se está analizando y X_i y β_i para $i = 1, ..., r$ son las variables de estudio y sus coeficientes respectivamente. Esta formulación es análoga a la que vimos en el capítulo anterior.

Como es habitual, el supuesto de riesgos proporcionales debe evaluarse para cualquier variable independiente del tiempo y en caso de que una o más de estas variables no cumplieran con este supuesto se necesitaría usar un modelo de Cox estratificado o un modelo de Cox extendido. También se utilizará este último si se trabajan con variables inherentemente dependientes del tiempo.

La principal diferencia en la forma en que se usa el modelo de Cox para analizar los datos de eventos recurrentes comparado con los de eventos no recurrentes (donde solo tenemos un intervalo de tiempo por sujeto) es la forma en que se tratan varios intervalos de tiempo del mismo sujeto. Estos intervalos intervienen en la creación de la función de probabilidad maximizada para el modelo de Cox empleado.

Para simplificar el análisis se suele asumir que los datos involucran solo variables independientes del tiempo las cuales satisfacen el supuesto de riesgos proporcionales. Para los datos de sobrevivencia recurrentes, un sujeto con más de un intervalo de tiempo permanece en el conjunto de riesgo hasta su último intervalo, después del cual el sujeto se retira del conjunto de riesgo. Por el contrario, para los datos de eventos no recurrentes, cada sujeto se elimina del conjunto de riesgos en el momento de la falla o la censura.

Un detalle importante es que, a pesar de que tanto en el análisis de datos de eventos recurrentes como en el de no recurrentes las diferentes líneas de datos se tratan como independientes, estas son diferentes dependiendo de los datos que estemos analizando. Específicamente en el caso de eventos recurrentes las diferentes líneas de datos aportadas por el mismo sujeto se traten en el análisis como si fueran contribuciones independientes de diferentes sujetos cuando en realidad no lo son. Basándonos en lo anterior tiene sentido considerar en el análisis que diferentes intervalos aportados por el mismo sujeto representan observaciones correlacionadas. Una

técnica ampliamente utilizada para ajustar esta correlación se conoce como *estimación robusta*, también conocida como *estimación empírica*. Esta técnica implica esencialmente ajustar las varianzas estimadas de los coeficientes de regresión ($\widehat{\texttt{Var}}(\widehat{\beta}_k)$) obtenidos al ajustar un modelo que considere la especificación errónea de la estructura de correlación asumida.

En el enfoque de procesos de conteo la estructura de correlación asumida es la independencia por lo que el modelo de riesgos proporcionales de Cox que se ajusta asume que diferentes resultados del mismo individuo son independientes. El objetivo de la estimación robusta para este enfoque es obtener estimadores de la varianza que se ajusten a la correlación dentro de los sujetos cuando anteriormente no se suponía tal correlación. La suposición que se encuentra atrás de esto es similar a la de un modelo mixto, en donde se asume que la media está correctamente especificada por el predictor lineal y la correlación solo afecta a la varianza. Es importante señalar que los coeficientes de regresión estimados en sí mismos no se ajustan; sólo se ajustan las varianzas estimadas de estos coeficientes. El estimador robusto de la varianza estimado permite pruebas de hipótesis e intervalos de confianza sobre los parámetros del modelo que dan cuenta de la correlación dentro de los sujetos.

La expresión para el estimador robusto de la varianza es la siguiente:

$$\widehat{\mathbf{R}}(\beta) = \widehat{\texttt{Var}}(\widehat{\beta}_k)\ [\widehat{\mathbf{R}}'_S\widehat{\mathbf{R}}_S]\ \widehat{\texttt{Var}}(\widehat{\beta}_k),$$

donde $\widehat{\texttt{Var}}(\widehat{\beta}_k)$ es la matriz de varianzas y covarianzas estimadas y $\widehat{\mathbf{R}}_S$ es la matriz de residuales. Ambas matrices se obtienen de la estimación mediante máxima verosimilitud parcial del modelo de Cox que se ajusta [15].

El estimador robusto para datos de eventos recurrentes fue planteado por Lin y Wei (1989) como una extensión similar al "estimador de información tipo sandwich" propuesto por Zeger y Liang (1986) para modelos lineales generalizados.

La fórmula de estimación robusta descrita anteriormente se aplica al enfoque de procesos de conteo, así como a otros enfoques para analizar los datos de eventos recurrentes.

Modelos Extendidos de Cox

Siempre que los eventos recurrentes involucren diferentes categorías de enfermedades y/o el orden de los eventos desempeñe un papel importante al abordar la verdadera pregunta de investigación, las técnicas de sobrevivencia son siempre una mejor opción que las técnicas estadísticas tradicionales.

Durante las últimas décadas, se han propuesto varios enfoques alternativos para el análisis de datos de eventos recurrentes que presenten las características anteriores. Estos métodos propuestos implican el uso de modelos estratificados de Cox de riesgos proporcionales y se pueden clasificar como: *modelos con corrección de varianza* y *modelos de fragilidad (frailty models)*. La principal diferencia entre estos dos tipos de modelos es la forma en que tratan la correlación

dentro de las observaciones.

Modelos de Varianza Corregida

En los modelos de varianza corregida (a veces también denominados "modelos de varianza robusta"), la correlación entre individuos debido a la heterogeneidad se halla ajustando la matriz de varianzas y covarianzas utilizando un estimador de jackknife agrupado y la correlación debida a la dependencia de eventos se encuentra mediante la construcción de diferentes conjuntos de riesgos los cuales se basan en diferentes intervalos de riesgo. Veamos a continuación un esbozo de cómo se halla la varianza de los coeficientes en estos modelos.

Como se mencionó anteriormente, 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 en [27], proponen una forma de calcular los valores de jackknife para obtener la varianza de los coeficientes considerando valores individuales mediante iteración de Newton-Raphson:

a. Obtener los coeficientes estimados de $\widehat{\beta}$ con todas las observaciones:

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + \mathsf{U}\mathcal{I}^{-1}$$

donde $\mathcal{I}^{-1}$ es la inversa de la matriz de información de Fisher y U es la matriz de residuales de tamaño $n \times p$.

b. Retirar la observación i

$$\mathsf{U}_{(i)1\times p}\, \mathcal{I}^{-1} = \mathtt{D}_i = \Delta\widehat{\beta}_{(i)} = \widehat{\beta}_{(i)}^{(k+1)} - \widehat{\beta}_{(i)}^{(k)}$$

de tal manera que $\mathtt{D}_{(i)}$ representa la i-ésima fila de la matriz $\mathbf{D}_{n\times p}$, que es la matriz de cambio en $\widehat{\beta}$ si es removida la i-ésima observación, o sea:

$$\mathbf{D} = \begin{bmatrix} \widehat{\beta}_{1(1)}^{(k)} - \widehat{\beta}_{1(1)}^{(k-1)} & \widehat{\beta}_{2(1)}^{(k)} - \widehat{\beta}_{2(1)}^{(k-1)} & \cdots & \widehat{\beta}_{p(1)}^{(k)} - \widehat{\beta}_{p(1)}^{(k-1)} \\ \widehat{\beta}_{1(2)}^{(k)} - \widehat{\beta}_{1(2)}^{(k-1)} & \widehat{\beta}_{2(2)}^{(k)} - \widehat{\beta}_{2(2)}^{(k-1)} & \cdots & \widehat{\beta}_{p(2)}^{(k)} - \widehat{\beta}_{p(2)}^{(k-1)} \\ \vdots & \vdots & \cdots & \vdots \\ \widehat{\beta}_{1(n)}^{(k)} - \widehat{\beta}_{1(n)}^{(k-1)} & \widehat{\beta}_{2(n)}^{(k)} - \widehat{\beta}_{2(n)}^{(k-1)} & \cdots & \widehat{\beta}_{p(n)}^{(k)} - \widehat{\beta}_{p(n)}^{(k-1)} \end{bmatrix}_{n\times p.}$$

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

Esta varianza puede ser escrita como:

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

Con grupos correlacionados el estimador sándwich estaría dado por $\tilde{\mathbf{D}}'\tilde{\mathbf{D}}$, donde $\tilde{\mathbf{D}}_{m\times p} = \mathbf{B}_{m\times n}\mathbf{D}_{n\times p}$, 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 $\tilde{\mathbf{D}}_{m\times p} = \mathbf{B}\mathsf{U}\mathcal{I}^{-1}$, entonces el cálculo de la varianza corregida para datos correlacionados, está dada por la siguiente expresión:

$$\tilde{\mathbf{D}}'\tilde{\mathbf{D}} = \mathcal{I}^{-1}\mathsf{U}'\mathbf{B}'\mathbf{B}\mathsf{U}\mathcal{I}^{-1}.$$

En modelos con corrección de varianza se han discutido variedades de modelos en la literatura entre los que se encuentran el modelo de incrementos independientes de Andersen y Gill (AG), el modelo de Prentice, Williams y Peterson (pueden ser de tiempo total o de intervalos de tiempo) (PWP-CP / PWP-GT), y el modelo marginal de Wei, Lin y Weissfeld (WLW). A continuación describiremos brevemente estos modelos.

-Modelo de incrementos independientes de Andersen-Gil

El *modelo de Andersen y Gill* (AG) es la extensión más simple de la regresión de riesgo proporcional de Cox utilizando el intervalo de tiempo del proceso de conteo [2]. Este asume que los eventos recurrentes dentro de los sujetos son independientes y que comparten un riesgo común de referencia. Siempre que sean válidas las suposiciones anteriores, este modelo proporcionará estimaciones más eficientes del coeficiente de regresión que la regresión de riesgo proporcional de Cox tradicional.

Debido a sus características este modelo resulta adecuado cuando las correlaciones entre los eventos para cada individuo son inducidas por covariables medidas y cuando un riesgo común de referencia para eventos repetidos puede justificarse teórica y estadísticamente.

Para el proceso de conteo de la entrada de datos en un modelo AG cada sujeto se representa como una serie de observación con tiempo de recurrencia dado como $(t_0, t_1]$, $(t_1, t_2]$, ..., $(t_{m-1}, t_m]$ (t_i tiempos de seguimiento) donde cada evento recurrente para el $i-$ésimo sujeto sigue un modelo de riesgo proporcional dado por

$$\lambda_i(t) = \lambda_0(t)\ \exp\{\beta_0\ \mathbf{Z}_i'(t)\},$$

donde $\lambda_0(t)$ es una función arbitraria de riesgo de referencia, β_0 es un vector de parámetros regresión de dimensión $M \times 1$ y el vector de covariables $\mathbf{Z}(t)$ se define como dependiente del tiempo para reflejar las influencias de eventos anteriores sobre las recurrencias futuras.

Según este modelo, el riesgo de un evento recurrente para un sujeto sigue el supuesto habitual de razón proporcional de Cox, pero no se tienen en cuenta el número de recurrencias. Los intervalos de riesgo de cada sujeto contribuyen al riesgo establecido para cada evento, independientemente del número de eventos para cada individuo.

Basándonos en la función de verosimilitud parcial de Cox podemos encontrar la logverosimilitud parcial de este modelo, la cual viene dada por la siguiente expresión:

$$\log L_p = \sum_{i=1}^{n} \int_0^t \mathbf{Z}_i'(u)\beta \mathrm{d}N_i(u) - \int_0^t \log\left\{\sum_{i=1}^{n} Y_i(u)\exp\left[\mathbf{Z}_i'(u)\beta\right]\right\}\mathrm{d}\overline{N}(u),$$

donde $N_i(t)$, en el contexto de eventos repetidos, denota al número de eventos en el tiempo t para un individuo i y $\overline{N} = \sum_{i=1}^{n} N_i$ representa al conjunto de riesgo para la ocurrencia de un

nuevo evento. Por su parte $Y_i(t)$ representa a un proceso predecible que toma valores en $\{0, 1\}$, donde 1 indica que el i-ésimo individuo está bajo observación y 0 que no lo está.

A diferencia del modelo clásico de Cox (1972) donde el individuo deja de estar en riesgo cuando $Y_i(t)$ toma el valor de cero, lo que quiere decir que el evento ocurrió, en el modelo propuesto por Andersen y Gill el valor que toma $Y_i(t)$ seguirá siendo uno si ocurre el evento dado que el individuo sigue en riesgo.

Técnicamente, el modelo de AG no difiere significativamente de un modelo de Cox estándar si cada evento repetido se ve como un episodio condicionalmente independiente dadas las covariables. Este modelo muestra también similitud con la regresión de Poisson ya que ambos se basan en el supuesto de incrementos independientes. El modelo AG tiene algunas ventajas sobre la regresión de Poisson, por ejemplo, mientras que la regresión de Poisson solo se puede usar para un riesgo uniforme a lo largo del tiempo, el modelo AG se puede usar también para un riesgo no sólo constante sino también proporcional. Otra ventaja del modelo AG es que, al ser un enfoque de sobrevivencia, utiliza más información (evento y tiempo del evento), lo que hace que el modelo AG en ciertas ocasiones aborde la pregunta de investigación de manera más apropiada que la regresión de Poisson.

Una dificultad del modelo de Andersen-Gill es que este no aborda particularmente la dependencia potencial de eventos repetidos dentro de los individuos, asumiendo que el evento recurrente de un individuo es condicionalmente independiente de eventos anteriores en presencia de ciertas covariables dependientes del tiempo basadas en la teoría. Bajo ciertas circunstancias tal especificación de covariables dependientes del tiempo, como el número de recurrencias previas e interacciones dadas, puede mitigar considerablemente la correlación intrapersonal, especialmente cuando se analizan datos de encuestas a gran escala. Sin embargo, empíricamente, la hipótesis de independencia no puede verificarse por medio de ningún medio estadístico y, por lo tanto, el supuesto no puede contrastarse estadísticamente. En muchas ocasiones cuando se trabaja con eventos repetidos, particularmente en el contexto de ensayos clínicos, dicha dependencia no puede ignorarse y la existencia de tal factor latente puede no explicarse completamente por los efectos de las covariables medibles. Dado esto, la fuerte suposición sobre la independencia condicional de eventos repetidos no siempre se ajusta a la correlación de recurrencias de eventos para el mismo individuo, incluso cuando se consideran covariables dependientes del tiempo. Como consecuencia, a menudo se considera necesario realizar ajustes estadísticos sobre el supuesto de independencia o utilizar alguna medida correctiva, como la corrección robusta de grupo.

A pesar de las desventajas que tiene el modelo de Andersen-Gill este proporciona una enorme simplicidad y conveniencia analítica para el modelado de eventos repetidos y su aplicación suele ser sugerida cuando existen un riesgo de referencia común para los eventos y este pueda justificarse teórica y estadísticamente.

-Modelos de Prentice Williams & Peterson

En 1981, Prentice, Williams y Peterson propusieron dos modelos para el análisis de eventos recurrentes [22], los cuales asumen que estos eventos dentro del sujeto están relacionados y el

riesgo de referencia varía de un evento a otro, diferenciándose así del modelo visto anteriormente. Por ejemplo, un evento que se podría modelar con estos sería el riesgo inicial de un segundo ataque cardíaco, ya que sabemos que este es siempre más alto que el riesgo inicial de un primer ataque cardíaco.

Estos dos modelos de PWP son muy semejantes y una de sus diferencias principales es que el primero se basa en el proceso de conteo y es conocido como *modelo de proceso de conteo de Prentice, Williams y Peterson* (PWP-CP), mientras que el segundo se basa en el intervalo de tiempo y se conoce como *modelo de intervalos de tiempo de Prentice, Williams y Peterson* (PWP-GT).

La característica del análisis de datos en la que el riesgo de referencia varía de un evento a otro y la ocurrencia de un evento posterior se ve afectada por un evento anterior está muy bien incorporada en ambos modelos de PWP. Ambos enfoques son modelos condicionales, ya que un sujeto no está en riesgo de su m-ésimo evento hasta que experimentó su (m-1) evento en el momento t, o sea, que el conjunto de riesgo del segundo evento considera solo aquellos sujetos que ya han experimentado su primer evento en el momento t.

Sea $N_i(t) = \{q(s) : s \leq t\}$ el número de eventos para el individuo i en el tiempo t, correspondiente a los tiempos de sobrevivencia aleatorios $T_{i1} < ... < T_{iQ}$, y $\mathbf{Z}_i(t)$ el vector de covariables de los individuos al tiempo t. Para un individuo con Q eventos repetidos antes de ser censurado, $t_0 = 0$, t_q el q-ésimo tiempo de evento recurrente para $q = 1, ..., Q$, y t_{Q+1} el tiempo censurado, tendremos que los modelos denominados PWP-CP y PWP-GT se definen mediante las funciones de riesgo siguientes:

$$\text{PWP-CP}: \quad \lambda_{iq}(t) = \lambda_{0q}(t) \exp\{\beta_q \, \mathbf{Z}'(t)\}$$

$$\text{PWP-GT}: \quad \lambda_{iq}(t) = \lambda_{0q}(t - t_{q-1}) \exp\{\beta_q \, \mathbf{Z}'(t)\}$$

donde $\lambda_{0q}(t)$ representa el riesgo de referencia específico para el k-ésimo evento a lo largo del tiempo.

Nótese que el modelo PWP-CP es similar al modelo AG-CP en el sentido en que un individuo se mueve al estrato q inmediatamente después de la ocurrencia del (q-1)-ésimo evento recurrente y permanece allí hasta la ocurrencia del q-ésimo evento o de la censura a la derecha. Estos dos enfoques se diferencian principalmente en que el modelo PWP-CP está estratificado por eventos por lo que los riesgos de referencia varían de un evento a otro, no así en el modelo AG-CP.

Aunque los modelos de tiempo total y tiempo de intervalo parecen similares en ciertos aspectos, existen importantes diferencias conceptuales entre estas dos perspectivas las cuales se deben a las diferentes especificaciones de los tiempos de eventos para un evento repetido. En el modelo PWP-CP, la función de intensidad de referencia depende del tiempo total t, incluso para las recurrencias posteriores. Por otro lado, el modelo PWP-GT describe un proceso de intensidad a partir de la ocurrencia de un evento inmediatamente anterior, con el tiempo de intervalo definido como $(t - t_{q-1})$; por lo tanto, cada proceso de sobrevivencia define un orden de clasificación diferente de un conjunto de riesgos específico sobre la ocurrencia de un evento

repetido. A pesar de sus diferencias ambos enfoques son clasificados como modelos condicionales ya que un individuo no se considera en el conjunto de riesgos para el q-ésimo evento hasta que experimenta el (q-1)-ésimo evento.

Las funciones de verosimilitud para estos modelos, al igual que en el caso analizado anteriormente, se pueden hallar utilizando el enfoque de la verosimilitud parcial de Cox. Sean $t_{q0} < t_{q1} < ... < t_{qd}$ los tiempos de sobrevivencia ordenados para el evento repetido q, d el número total de eventos y $\mathbf{Z}_{iq}(t_{iq})$ el vector de covariables. Tendremos que la función de verosimilitud parcial viene dada por:

$$L_p(\beta_q) = \prod_{q\geq 1}\prod_{i=1}^{d_q} \frac{\exp\{\beta_q \mathbf{Z}'_{qi}(t_{qi})\}}{\sum_{l\in R(t_{qi},q)} \exp\{\beta_q \mathbf{Z}'_l(t_{qi})\}}$$

Como $t_{q0} = 0$, la verosimilitud parcial anterior le corresponde al modelo PWP-CP, la cual describe el proceso de sobrevivencia completo para la ocurrencia de una serie de eventos repetidos.

Modificando la anterior y teniendo en cuenta que la estratificación en el modelo PWP-GT está restringida de manera que un individuo pueda contribuir como máximo en un evento en un estrato específico, podemos obtener la función de verosimilitud para el modelo PWP-GT. Sea $g_{q1} < ... < g_{q_{e_q}}$ los distintos intervalos de tiempos ordenados a partir de la falla inmediatamente anterior y e_q el número total de fallas que ocurren en el estrato q. Definamos también el conjunto de riesgo $R(g, q)$ para el modelo PWP-GT como un conjunto que contenga a aquellos que han experimentado el (q-1)-ésimo evento con $t_{q0} = t_{q-1}$. Dadas las condiciones anteriores tendremos que:

$$L_p(\beta_q) = \prod_{q\geq 1}\prod_{i=1}^{e_q} \frac{\exp\{\mathbf{Z}'_{qi}(t_{qi})\beta_q\}}{\sum_{l\in R(t_{q_i,q})} \exp\{\mathbf{Z}'_l(\widehat{l}_l + g_{qi})\beta_q\}}$$

donde $\widehat{l}_l$ representa el último tiempo en que falló el individuo l antes de entrar al evento q.

Nótese que en las expresiones de las verosimilitudes parciales para estos dos modelos el numerador es el mismo por lo que su diferencia reside en el conjunto de riesgos y en cómo se define este.

El modelo PWP-CP puede usarse si uno está interesado en conocer el efecto de la intervención sobre la variable de resultado desde el comienzo del estudio, mientras que PWP-GT debe usarse si uno está interesado en conocer el efecto de eventos previos.

Los modelos PWP tienen una ventaja adicional sobre otros modelos ya que como tienen un riesgo de referencia específico del evento, estos pueden estimar un efecto general o un efecto específico del evento para cada covariable. Aunque ambos modelos son muy apropiados para el análisis de eventos recurrentes, tienen algunas limitaciones, por ejemplo, estos pueden proporcionar estimaciones poco fiables para eventos de orden superior debido a que a medida que

aumenta el orden de eventos disminuye el número de sujetos en el conjunto de riesgo.

-Modelo marginal de Wei, Lin & Weissfeld

Wei, Lin y Weissfeld en 1989 propusieron un modelo tipo Cox para analizar datos de eventos repetidos. El *modelo de Wei, Lin y Weissfeld* [30] es el único modelo con corrección de varianza que se puede aplicar tanto a múltiples fallas del mismo tipo de eventos como a múltiples fallas de diferentes tipos de eventos. En este se considera la recurrencia de cada evento como un proceso separado y no hay ningún orden entre los eventos dentro del sujeto.

Por ejemplo, durante la estadía en la unidad de cuidados intensivos neonatales, un recién nacido corre el riesgo de sufrir varios eventos simultáneamente, como infección por microorganismos, enterocolitis necrotizante, meningitis, ictericia y diarrea, etc. Cada uno de estos eventos puede ocurrir más de una vez en cualquier orden. El modelo WLW analiza simultáneamente el tiempo hasta la detección del primer, segundo, tercer o más incidentes de varios tipos de eventos, ya sea en la misma visita clínica o en una diferente. El conjunto de riesgo establecido para el m-ésimo evento en el modelo WLW incluye a todas las personas que aún no han experimentado su m-ésimo evento y permanecen en seguimiento en el momento t. Por ejemplo, el conjunto de riesgos para el segundo evento incluiría a todos los individuos que no han experimentado sus segundos eventos y permanecieron en el seguimiento en el momento t; en otras palabras, el conjunto de riesgos incluye aquellos que no han experimentado ningún evento y están en seguimiento en el momento t y a aquellos que han experimentado un solo evento y están en seguimiento en el momento t.

La función de riesgo para el q-ésimo evento para el i-ésimo individuo es:

$$\lambda_{iq}(t) = Y_{iq}(t)\lambda_{0q}(t)\ \exp\{\beta_q \mathbf{Z}_i'(t)\}$$

y la función de verosimilitud parcial para el q-ésimo evento específico esta dada por:

$$L_q = \prod_{i=1}^{n} \left(\frac{\exp\{\beta\ \mathbf{Z}_{iq}(T_{iq})\}}{\sum_{l \in R_q(T_{iq})} \exp\{\beta\ \mathbf{Z}_{iq}(T_{iq})\}} \right)^{\delta_{iq}}$$

donde $\lambda_{0k}(t)$ es una función de riesgo base no especificada, Y_{ik} es el indicador de riesgo para el modelo y $R_k(t) = \{l : T_{kl} \geq t\}$ es el conjunto de sujetos que están en riesgo en el k-ésimo evento previo al tiempo t.

A diferencia del modelo de Andersen-Gill, este modelo permite una probabilidad por separado para cada evento y cuando ésta es cero significa que el sujeto ya no está en riesgo después del último evento dado. Este modelo proporciona estimaciones confiables del coeficiente de regresión cuando los datos no siguen un orden específico ya que si el orden importa, este exagera el verdadero efecto porque permite que un sujeto esté en riesgo durante varias veces por un mismo evento. Debido a lo anterior se ha criticado el uso de este modelo en el campo del análisis de eventos recurrentes ya que, en general, estos siguen un orden; no sucede así con el análisis de eventos competitivos donde estos modelos suelen ser más aceptados y útiles.

Modelos de fragilidad

Los *modelos de fragilidad* son otra clase de modelos extendidos a partir del modelo de riesgo proporcional tradicional de Cox [28]. A diferencia de los modelos con corrección de varianza vistos anteriormente, estos modelos asumen que la correlación entre eventos recurrentes se debe a la tendencia de que algunos individuos sean más propensos a desarrollar eventos recurrentes en comparación con otros debido a algunos factores no observados o desconocidos, los cuales pueden ser por ejemplo factores sociodemográficos, ambientales, de comportamiento o genéticos [31]. Muchas veces estos factores son desconocidos para el investigador y difíciles de incorporar al análisis.

Para corregir la estimación de los parámetros de regresión, vimos que en los casos de modelos de varianza corregida se realizaba un ajuste a la matriz de varianzas y covarianza, en estos nuevos modelos se incorpora un término de fragilidad directamente a la estimación del modelo, dicho término sigue una distribución de fragilidad específica.

La función de riesgo para el tiempo recurrente del k-ésimo evento en el i-ésimo sujeto ($k = 1, 2, ..., k_i \ ; \ i = 1, 2, ..., n$) condicionado a la fragilidad Z_i, sigue la forma de riesgos proporcionales y está dado por:

$$\lambda_{ik}(t) = \lambda_{0k}(t) Z_i \exp\{x_i(t)\beta_k\}, \quad t > 0$$

donde, $\lambda_{0k}(t)$ es la función de riesgo de referencia, X_i es un vector de covariables observables y β es un vector desconocido de coeficientes de regresión. La fragilidad se denota con el término Z_i y son los factores de riesgo comunes no observados (aleatorios) compartidos por todos los sujetos del grupo i y se supone que es una variable aleatoria independiente igualmente distribuida con media unitaria y varianza desconocida θ.

En el caso en que pudiéramos identificar correctamente la distribución de la fragilidad, estos modelos serían más eficientes que los modelos con corrección de la varianza. Actualmente no hay pautas sobre cómo seleccionar la distribución de fragilidad adecuada para cada caso, en general suele emplearse con mayor frecuencia la distribución Gamma para estimar el término de fragilidad debido a sus características y popularidad. Otras distribuciones que se utilizan para la estimación de la fragilidad son la distribución normal, la distribución logarítmica normal y la distribución uniforme.

-Modelo estándar de fragilidad

El *modelo de fragilidad estándar* es la extensión que involucra términos de fragilidad más simple del modelo de riesgo proporcional de Cox y es muy similar al modelo Andersen-Gill. Al igual que el modelo AG, asume que no hay corrección dentro de los sujetos debido a la dependencia del evento y que cualquier correlación que esté presente entre los eventos recurrentes se debe solo a la heterogeneidad. Similar al modelo AG, el riesgo de referencia se asume común para todos los eventos en el modelo de fragilidad estándar. Una diferencia entre estos modelos y los modelos AG es que en estos el término de fragilidad se incorpora directamente en el modelo y la ecuación paramétrica estructural se usa para estimar el término de fragilidad

para la estimación de la correlación dentro del sujeto mientras que en el caso del modelo AG dentro del sujeto la correlación se halla ajustando la matriz de varianzas y covarianzas.

El modelo de fragilidad estándar es computacionalmente muy intenso y requiere mucho más tiempo que el modelo AG y su interpretación tampoco es tan sencilla. Generalmente, el modelo de fragilidad se interpreta como mantener constante el término de fragilidad entre los individuos, lo que intuitivamente no es aceptable para muchos investigadores.

-Modelo condicional de fragilidad

Muchas veces es difícil distinguir entre las fuentes de correlación dentro del sujeto, es decir, si se debe a la dependencia del evento, a la heterogeneidad o a ambas. En vista de esto, se agregó el término de fragilidad en el modelo PWP-GT para que dentro del sujeto la correlación debida a cualquiera de las fuentes pudiera acomodarse en el modelo. Este nuevo modelo se conoce como *modelo de fragilidad condicional*. Básicamente, la idea de este es que, dentro de la correlación del sujeto debido a la dependencia del evento se acomodará la naturaleza condicional del modelo, es decir, un sujeto no está en riesgo de un evento m-ésimo hasta que experimente su evento (m-1)-ésimo y entre los sujetos la correlación debida a la heterogeneidad se adaptará al incorporar el término de fragilidad en el proceso de estimación del modelo en sí.

Este modelo es relativamente más nuevo y hasta ahora pocos investigadores han trabajado en él, por lo tanto, se necesita más investigación en este con el fin de conocer su eficacia en datos reales [31].

Ajuste y selección de los modelos

En el caso de los modelos presentados en esta sección y que analizan eventos recurrentes pero no están basados en MLG, se utilizan diferentes pruebas de bondad de ajustes. Estas son las mismas que para los modelos basados en MLG, vistas y analizadas anteriormente, solo que en este caso debemos adaptarlas en cierto sentido a las nuevas especificaciones de cada modelo. En estos modelos también resulta de interés el análisis de los residuos para concluir sobre el ajuste en general.

El análisis residual y la medida de influencia de cada observación generalmente se realiza usando el valor de jackknife agrupado que evalúa cada punto en el ajuste del modelo [27]. Este se define como:

$$\mathbf{j}_i = \widehat{\beta_{(i)}} - \widehat{\beta}$$

donde $\widehat{\beta}_{(i)}$ es el resultado de un ajuste el cual incluiría todos los individuos excepto al individuo i.

El residuo jackknife se puede calcular entonces usando el mismo esquema iterativo del algoritmo para hallar las varianzas corregidas, de tal manera que la matriz de valores influyentes jackknife estaría conformada por:

$$\mathbf{J} = \begin{bmatrix} \widehat{\beta}_{1(1)} - \widehat{\beta} & \widehat{\beta}_{2(1)} - \widehat{\beta} & \dots & \widehat{\beta}_{p(1)} - \widehat{\beta} \\ \widehat{\beta}_{1(2)} - \widehat{\beta} & \widehat{\beta}_{2(2)} - \widehat{\beta} & \dots & \widehat{\beta}_{p(2)} - \widehat{\beta} \\ \vdots & \vdots & \dots & \vdots \\ \widehat{\beta}_{1(n)} - \widehat{\beta} & \widehat{\beta}_{2(n)} - \widehat{\beta} & \dots & \widehat{\beta}_{p(n)} - \widehat{\beta} \end{bmatrix}_{n \times p}$$

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 en el caso que estamos estudiando sería proceder mediante Newton-Raphson para el modelo de Cox

$$\Delta\beta = \mathbf{1}'(\mathsf{U}\mathcal{I}^{-1}) = \mathbf{1}'\mathbf{D}$$

donde la matriz $\mathbf{D}$ es llamada matriz de residuales de β, y U puede ser obtenida de la siguiente forma:

$$\mathsf{U} = \begin{bmatrix} \dfrac{\partial L}{\partial \beta_{1(1)}} & \dfrac{\partial L}{\partial \beta_{2(1)}} & \dots & \dfrac{\partial L}{\partial \beta_{p(1)}} \\ \dfrac{\partial L}{\partial \beta_{1(2)}} & \dfrac{\partial L}{\partial \beta_{2(2)}} & \dots & \dfrac{\partial L}{\partial \beta_{p(2)}} \\ \vdots & \vdots & \dots & \vdots \\ \dfrac{\partial L}{\partial \beta_{1(n)}} & \dfrac{\partial L}{\partial \beta_{2(n)}} & \dots & \dfrac{\partial L}{\partial \beta_{p(n)}} \end{bmatrix}_{n \times p}$$

Estimación e interpretación de los parámetros

Antes de concluir este capítulo, quisieramos destacar algunos aspectos generales sobre la estimación e interpretación de los parámetros de los modelos anteriores.

- En el caso de los modelos con varianza corregida estudiados (AG, PWP-CP, PWP-GT, WLW) para estimar el vector de parámetros $\widehat{\beta}$ se utiliza la verosimilitud parcial propuesta por D. R. Cox en 1975. Específicamente, la estimación vendrá dada por la solución de la ecuación que se obtiene al derivar el logaritmo de la verosimilitud parcial en el tiempo t e igualar a cero. Al igual que en los modelos basados en MLG, esta solución suele obtenerse por métodos iterativos mediante el algoritmo de Newton-Raphson.

  Las estimaciones que se obtienen serán estimaciones robustas y consistentes de estos parámetros siempre y cuando los modelos estén bien especificados. Posteriormente se llevan a cabo pruebas como la conocida prueba robusta de Wald y la prueba robusta de Score para verificar estos resultados.

  - Prueba robusta de Wald: Bajo la recurrencia del evento de interés, las pruebas usuales se consideran no conservadoras y una de las alternativas existentes es esta prueba.

Esta se basa en la estimación $\widehat{\beta}'[\tilde{D}'\tilde{D}]^{-1}\widehat{\beta}$.

- Prueba robusta de Score: El estadístico para esta prueba está basado en la primera iteración para estimar $\widehat{\beta}$ con el algoritmo de Newton-Raphson, y viene dado por

$$S = [\mathbf{1}'\mathsf{U}]\mathcal{I}^{-1}[\mathsf{U}'\mathbf{1}]$$

Al sustitur la inversa del estimador de la varianza, recordando que $\mathcal{I} = \mathsf{U}'\mathsf{U}$ y considerando la variable indicadora como corrección $\tilde{\mathcal{I}} = \mathsf{U}'\mathbf{B}'\mathbf{B}\mathsf{U}$ se obtiene el estadístico de la prueba de score:

$$S_r = [\mathbf{1}'\mathsf{U}][\mathsf{U}'\mathbf{B}'\mathbf{B}\mathsf{U}]^{-1}[\mathsf{U}'\mathbf{1}]$$

donde la primera iteración es $\widehat{\beta}^{(1)} = \widehat{\beta}^{(0)} + \mathsf{U}\mathcal{I}^{-1}$ y U y $\mathcal{I}^{-1}$ se calculan a partir de $\widehat{\beta}^{(0)}$.

Ambas pruebas anteriores tienen como hipótesis nula $H_0 : \beta = \beta^{(0)}$ y la distribución de cada uno de los estadísticos para probar dicha hipótesis se distribuye como una ji-cuadrada con p grados de libertad.

- En el caso de los modelos con fragilidades un algoritmo bastante usado en la estimación de los parámetros del modelo es el algoritmo EM. Este tiene la desventaja de realizar las estimaciones de los coeficientes de manera lenta y con un gran número de operaciones computacionales en la mayoría de los casos.

  Debido a lo anterior existen diferentes vías alternativas, una de ellas son los modelos penalizados en los cuales los términos de fragilidad son tratados como coeficientes de regresión adicionales, los cuales están limitados por una función de penalización que se agrega a la log-verosimilitud [28]. Sin embargo, Rondeau y Gonzalez, mencionan que esos métodos, al igual que el algoritmo EM, presentan inconvenientes debido a la lenta convergencia para estimar los parámetros, asimismo no estiman la varianza de la fragilidad y no pueden ser usados para estimar la función de riesgo [24].

  Otra alternativa sería modificar el algoritmo EM, donde dicha modificación sería en el sentido de introducirle un nuevo paso con el fin de modelizar la varianza de la fragilidad directamente. Un ejemplo del algoritmo EM modificado, denotado por EMB puede verse en [3].

Finalmente, ya que tenemos un modelo adecuado a nuestros datos y nuestras preguntas de investigación, el proceso de modelado culmina con la interpretación del modelo. La interpretación de los parámetros del modelo se realiza al igual que en los modelos basados en MLG. Como la ecuación del modelo está expresada en términos multiplicativos, los parámetros se interpretan en términos del factor de cambio en el valor esperado para un incremento unitario de las variables explicativas, o sea, se tendría que para un cambio de δ unidades en la variable explicativa x_j manteniendo el resto de variables constantes, el recuento esperado se incrementa en un factor de $\exp\{\beta_j\ \delta\}$.

Capítulo 3

Aplicaciones

En la presente sección aplicaremos los métodos de análisis de eventos recurrentes a cuatro conjuntos de datos simulados con diferentes características para comparar el comportamiento de los modelos y sus resultados. Posteriormente, aplicaremos los conceptos de análisis de sobrevivencia y de análisis de eventos recurrentes a una base de datos reales con el objetivo de plantear modelos de análisis de eventos recurrentes para la misma e interpretar los resultados que se obtengan.

3.1. Simulación

En esta sección, como mencionamos al inicio de este capítulo, simularemos datos con ciertas características con el objetivo de aplicarle los modelos para eventos recurrentes y obtener en cada caso las estimaciones de sus parámetros. Estas estimaciones las compararemos con los valores predefinidos y analizaremos el comportamiento de cada modelo dependiendo de los rasgos distintivos de cada conjunto de datos.

3.1.1. Datos simulados

Se simularon cuatro tipos de bases de datos de eventos recurrentes, con las características comunes siguientes:

- Todas tienen solo una covariable.
- El tiempo de seguimiento está comprendido entre 0 y 10.
- En estas tendremos las variables:

  id: Identificador de cada sujeto. (Se encuentra repetido para cada recurrencia)

  enum: Número del evento.

  t.start: Inicio del intervalo. (Toma el valor 0 o el tiempo de recurrencia anterior)

  t.stop: Tiempo de censura o tiempo en el que ocurre el evento nuevamente.

event: Variable dicotómica que indica si sucedió (event = 1) o no (event = 0) el evento de interés.

X: Covariable considerada. Su distribución varía dependiendo de la base de datos que estemos considerando.

- Para cada una de las anteriores se consideraron 3 variaciones dependiendo de la cantidad de individuos incluidos en el estudio ($n \in \{20, 30, 50\}$).

- Para simular las bases de datos se emplearon las librerías "reReg" [25] y "reda" [29] de R [23].

A pesar de que como vimos las bases de datos generadas tiene varias características en común estas también poseen rasgos distintivos que las diferencian entre sí, estos son:

BD1- Los datos se generaron sin considerar ningún término frailty y la covariable se distribuye como números aleatorios independientes provenientes de una distribución normal estándar, por lo tanto podemos afirmar que esta no depende del tiempo. El coeficiente de la variable **X** se tomó igual a 1.3 sin importar el número de individuos que se estuviera analizando. No se consideró censura de los datos, por lo que en los tres casos analizados todos los individuos presentan eventos hasta el tiempo final.

BD2- Nuevamente generamos los datos sin considerar ningún término frailty pero en este caso la covariable **X** se genera dependiente del tiempo t, mediante la siguiente función $\mathbf{X}_t = f(t) = t * u/5$ donde u es un valor aleatorio perteneciente a la distribución normal estándar. El coeficiente de la variable **X** se tomó igual a -3.85, -2.9 o -1.95 dependiendo de si estabamos trabajando con 20, 30 o 50 individuos respectivamente. Al igual que en el caso anterior la probabilidad de obtener datos censurados se fijó en 0 por lo que los datos para cada uno de los individuos del estudio estarán completos.

BD3- En este caso los datos sí se generaron considerando un término frailty tomado como un valor aleatorio de una distribución gamma con media 1 y varianza 0.25. La covariable **X** no depende del tiempo y se tomó nuevamente como valores aleatorios independientes provenientes de una distribución normal estándar. El coeficiente de la covariable se tomó igual a -1 para los 3 casos analizados. Para estos tres conjuntos sí admitimos la posibilidad de censura la cual se generaba a partir de una distribución aleatoria uniforme.

BD4- Para estos datos también consideramos un término frailty el cual se tomó igual a 0.8. La covariable **X**, al igual que en el caso 2 se definió dependiente del tiempo mediante $\mathbf{X}_t = f(t) = t * u^2/5$. Los coeficientes de la covariable **X** se tomaron igual a -3, -4 y -5 dependiendo de si era el estudio de 20, 30 o 50 individuos. Se tomó la probabilidad de censura una vez más igual a 0.

Importante mencionar que en cada caso las bases de datos se generaron 100 veces y los resultados que se reportan, a no ser que se especifique lo contrario, corresponden a los valores medios.

A continuación veremos tablas que resumen los valores medios de las principales características de las variables para cada conjunto de bases de datos. También realizamos gráficos los cuales incluyen los eventos recurrentes y los eventos finales de cada individuo para una realización particular con el objetivo de mostrar su distribución a través del tiempo.

BD1

Tabla 3.1: Bases de datos 1

N	Ident	Min	1st Qu.	Median	Mean	3er Qu.	Max	0's	1's
n=20									
	`enum`	1.000	8.825	26.70	42.072	65.908	152.7		
	`t.start`	0.000	2.174	4.791	4.782	7.370	9.973		
	`t.stop`	0.023	2.649	5.262	5.239	7.864	10.00		
	`X`	-1.870	0.591	1.256	1.090	1.807	1.908		
	`event`							20	490
n=30									
	`enum`	1.000	8.680	28.24	48.30	75.52	189.3		
	`t.start`	0.000	2.168	4.757	4.775	7.372	9.986		
	`t.stop`	0.016	2.629	5.230	5.223	7.857	10.00		
	`X`	-2.071	0.595	1.343	1.139	1.875	2.065		
	`event`							30	721
n=50									
	`enum`	1.000	8.455	28.65	56.46	84.95	253.5		
	`t.start`	0.000	2.170	4.800	4.792	7.388	9.991		
	`t.stop`	0.010	2.633	5.253	5.227	7.848	10.00		
	`X`	-2.306	0.563	1.326	1.180	1.944	2.276		
	`event`							50	1197

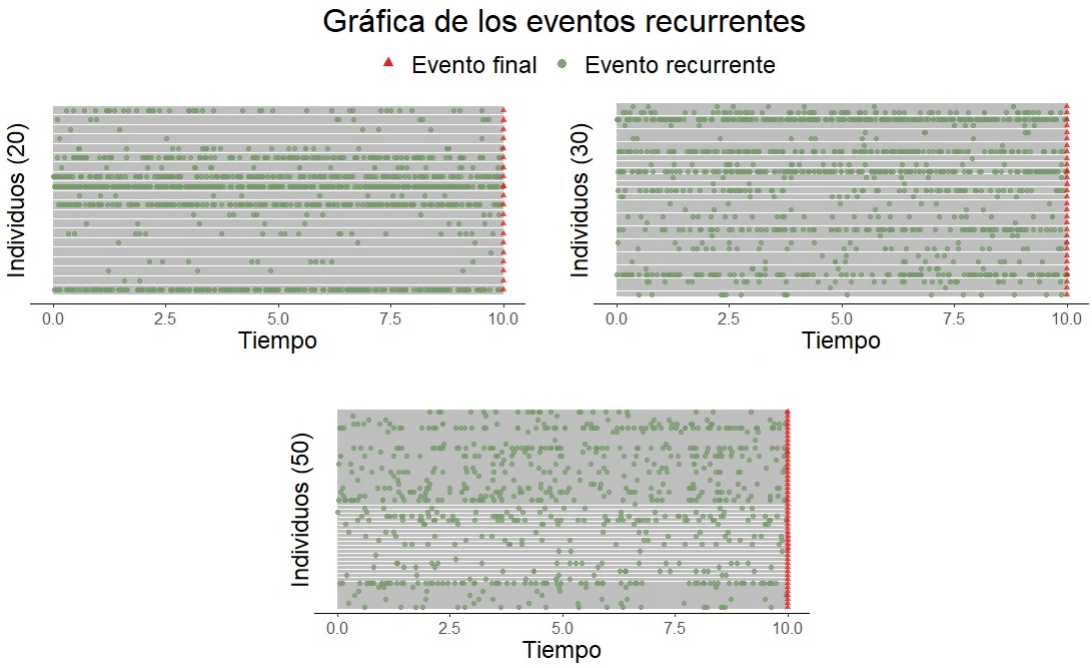

Figura 3.1: Eventos por individuos para Bases de Datos 1

BD2

Tabla 3.2: Bases de datos 2

N	Ident	Min	1st Qu.	Median	Mean	3er Qu.	Max	0's	1's
n=20									
	`enum`	1.000	1.023	2.080	2.676	3.513	8.610		
	`t.start`	0.000	0.004	1.104	2.273	3.917	9.295		
	`t.stop`	0.140	2.444	6.320	6.006	9.989	10.00		
	`X`	0.005	0.103	0.279	0.504	0.754	1.895		
	`event`							20	35
n=30									
	`enum`	1.000	1.050	2.260	2.981	3.968	10.36		
	`t.start`	0.000	0.006	1.722	2.707	4.682	9.751		
	`t.stop`	0.075	2.373	5.682	5.782	9.983	10.00		
	`X`	0.003	0.099	0.282	0.484	0.691	1.933		
	`event`							30	69
n=50									
	`enum`	1.000	1.840	2.910	3.430	4.640	12.20		
	`t.start`	0.000	0.183	2.494	3.178	5.411	9.906		
	`t.stop`	0.035	2.300	5.204	5.506	9.575	10.00		
	`X`	0.001	0.107	0.292	0.469	0.654	1.966		
	`event`							50	166

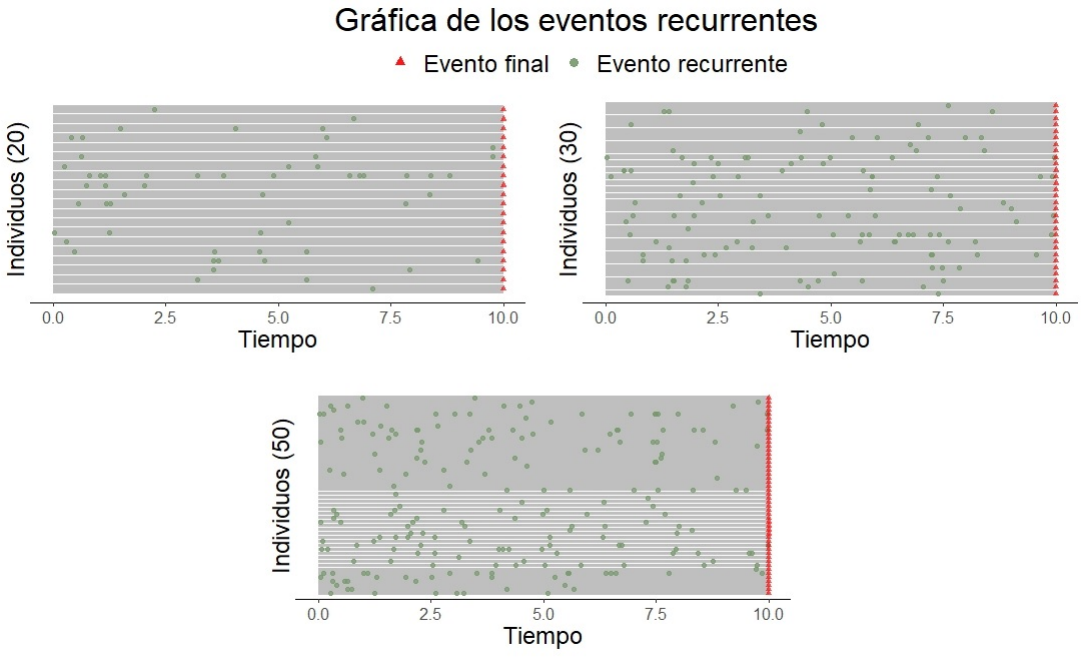

Figura 3.2: Eventos por individuos para Bases de Datos 2

BD3

Tabla 3.3: Bases de datos 3

N	Ident	Min	1st Qu.	Median	Mean	3er Qu.	Max	0's	1's
n=20									
	`enum`	1.000	2.580	6.36	10.00	15.03	36.37		
	`t.start`	0.000	0.240	1.172	2.063	3.164	9.465		
	`t.stop`	0.012	0.658	1.848	2.888	4.351	9.995		
	`X`	-1.898	-1.597	-1.040	-0.855	-0.286	1.797		
	`event`							20	121
n=30									
	`enum`	1.000	2.435	6.170	10.75	16.46	42.18		
	`t.start`	0.000	0.226	1.156	2.084	3.150	9.602		
	`t.stop`	0.010	0.640	1.847	2.921	4.389	10.00		
	`X`	-2.020	-1.569	-0.961	-0.818	-0.237	1.971		
	`event`							30	171
n=50									
	`enum`	1.000	2.413	6.035	11.20	15.39	53.17		
	`t.start`	0.000	0.221	1.132	2.094	3.189	9.821		
	`t.stop`	0.006	0.620	1.810	2.914	4.397	10.00		
	`X`	-2.255	-1.641	-0.962	-0.876	-0.253	2.219		
	`event`							50	284

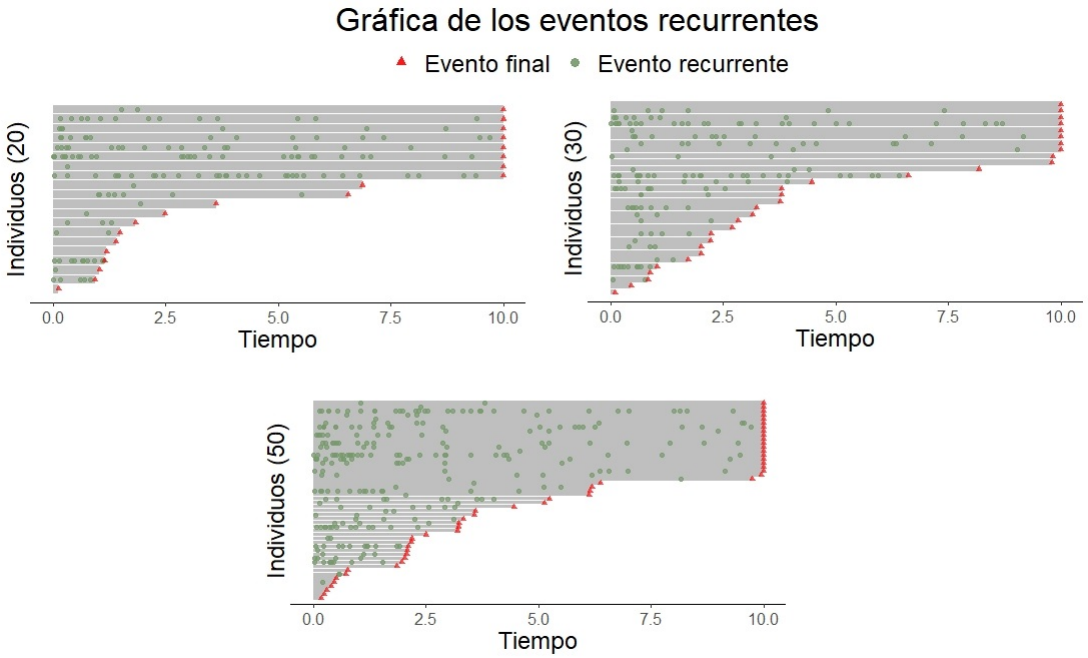

Figura 3.3: Eventos por individuos para Bases de Datos 3

BD4

Tabla 3.4: Bases de datos 4

N	Ident	Min	1st Qu.	Median	Mean	3er Qu.	Max	0's	1's
n=20									
	`enum`	1.000	1.303	2.480	3.154	4.283	9.790		
	`t.start`	0.000	0.068	2.162	3.022	5.238	9.740		
	`t.stop`	0.136	2.569	5.826	5.842	9.849	10.00		
	`X`	0.0005	0.033	0.127	0.294	0.358	1.827		
	`event`							20	53
n=30									
	`enum`	1.000	1.028	2.275	2.991	4.113	10.04		
	`t.start`	0.000	0.004	1.644	2.792	4.974	9.741		
	`t.stop`	0.085	2.515	6.358	5.999	9.989	10.00		
	`X`	0.0003	0.025	0.102	0.296	0.344	1.854		
	`event`							30	65
n=50									
	`enum`	1.000	1.000	2.140	3.004	4.073	11.32		
	`t.start`	0.000	0.000	1.462	2.676	4.795	9.838		
	`t.stop`	0.059	2.656	6.698	6.164	10.00	10.00		
	`X`	0.00009	0.022	0.092	0.293	0.325	1.916		
	`event`							50	95

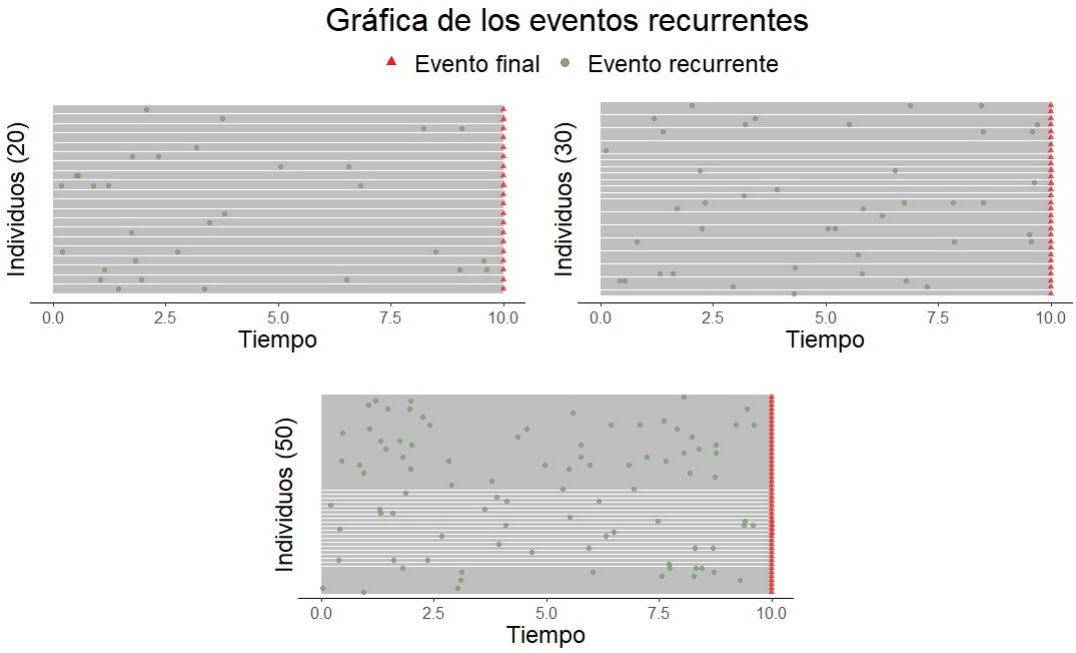

Figura 3.4: Eventos por individuos para Bases de Datos 4

3.1.2. Resultados

Una vez simuladas las 100 bases de datos para cada uno de los cuatro casos, ajustamos a cada una de ellas modelos de eventos recurrentes de Andersen-Gill, Prentice-Williams-Peterson (en sus dos versiones) y modelos de fragilidad, con el objetivo de comprobar qué modelo se ajustaba mejor en cada uno de los casos. A continuación se reportan los resultados, donde los valores que se muestran son la media de cada resultado obtenido, excepto para el valor de estadístico y su p-valor, para los cuales se seleccionó el máximo de los obtenidos, garantizando así que si el "modelo" mostrado es significativo, todos los modelos que se construyeron para ese caso también lo son.

BD1

En el primer caso, donde la covariable no era dependiente del tiempo y no había presencia de frailty en los datos, a pesar de que todos los modelos resultaron significativos para algún nivel, el que mejor ajustó los coeficientes a los valores reales fue el modelo de Andersen-Gil. En el caso de 20 individuos el estimado que se obtuvo fue de 1.307, para 30 individuos de 1.306 y para 50 individuos se obtuvo el valor 1.301, siendo en todos estos casos el valor real 1.3. A continuación tenemos un resúmen de los modelos y los resultados de cada uno.

Tabla 3.5: Media de las estimaciones de los modelos para las Bases de datos 1

N	Modelo	coef(X)	se(coef(X))	rob.se	L.95	U.95	Est.	p-valor	sd(coef)
n=20									
	AG	**1.307**	**0.068**	**0.058**	**1.196**	**1.425**	**55.06**	$\mathbf{2.9e-12}^{***}$	**0.068**
	PWP-CP	1.376	0.275	0.218	0.969	1.867	11.20	$1.8e-04^{***}$	0.273
	PWP-GT	1.310	0.134	0.115	1.092	1.550	23.79	$7.07e-05^{***}$	0.118
	Frailty	0.235	0.065	14.04	0.106	0.372	1.000	$0.078^{\cdot}$	0.079
n=30									
	AG	**1.306**	**0.051**	**0.045**	**1.220**	**1.40**	**90.04**	$\mathbf{< 2e-16}^{***}$	**0.045**
	PWP-CP	1.341	0.201	0.167	1.035	1.707	15.45	$7.5e-08^{***}$	0.221
	PWP-GT	1.314	0.102	0.089	1.145	1.494	25.61	$< 2e-16^{***}$	0.094
	Frailty +	0.220	0.049	21.03	0.124	0.322	1.000	$0.032^{\cdot}$	0.065
n=50									
	AG	**1.301**	**0.036**	**0.032**	**1.239**	**1.364**	**140.5**	$\mathbf{< 2e-16}^{***}$	**0.041**
	PWP-CP	1.339	0.143	0.128	1.096	1.604	14.92	$8.4e-13^{***}$	0.146
	PWP-GT	1.303	0.072	0.064	1.179	1.431	34.14	$< 2e-16^{***}$	0.066
	Frailty	0.207	0.035	33.33	0.136	0.281	1.000	$7.3e-04^{***}$	0.052

Signif. codes: '***' 0.001, '**' 0.01 '*' 0.05, '.' 0.1, ' ' 1
'Frailty +' modelos con término de fragilidad significativo

La última columna corresponde a las desviaciones estándar de las estimaciones de los coeficientes para cada modelo ajustado en cada caso. Podemos observar que las menores variaciones se obtienen para las estimaciones obtenidos con el modelo de Andersen-Gil.

BD2

Para el segundo grupo de bases de datos, nuevamente todos los modelos resultaron significativos pero el que tuvo mejor desempeño fue el modelo de PWP-CP. Lo anterior era lo esperado ya que estos datos se generaron considerando a la covariable dependiente del tiempo, característica que está tomada en cuenta en la formulación de los modelos de PWP. En el caso de 20 individuos el coeficiente estimado que se obtuvo fue de -3.975, para 30 individuos de -2.932 y para 50 individuos se obtuvo el valor -2.193, siendo los valores reales -3.85, -2.9 y -1.95 respectivamente. Se muestra en la siguiente tabla un resúmen de los modelos y de sus ajustes.

Tabla 3.6: Media de las estimaciones de los modelos para las Bases de datos 2

N	Modelo	coef(X)	se(coef(X))	rob.se	L.95	U.95	Est.	p-valor	sd(coef)
n=20									
	AG	-3.720	0.861	0.893	-4.649	-1.712	-2.686	0.007^{**}	1.061
	PWP-CP	**-3.975**	**1.178**	**1.010**	**-4.728**	**-1.697**	**-2.351**	$\mathbf{0.019}^{*}$	1.309
	PWP-GT	-4.007	1.118	0.976	-4.760	-1.808	-2.327	0.020^{*}	1.270
	Frailty	-6.743	1.299	21.72	-5.177	-2.320	1.000	0.005^{**}	4.318
n=30									
	AG	-2.990	0.609	0.570	-3.678	-1.736	-3.002	0.003^{**}	0.662
	PWP-CP	**-2.932**	**0.497**	**0.566**	**-3.710**	**-1.710**	**-3.153**	$\mathbf{0.002}^{**}$	0.766
	PWP-GT	-3.097	0.591	0.562	-3.802	-1.864	-3.243	0.001^{**}	0.731
	Frailty+	-4.809	0.696	40.45	-4.625	-2.559	1.000	$4.83e-05^{***}$	1.958
n=50									
	AG	-2.196	0.309	0.302	-2.709	-1.562	-4.722	$2.3e-06^{***}$	0.351
	PWP-CP	**-2.193**	**0.275**	**0.311**	**-2.718**	**-1.535**	**-4.726**	$\mathbf{2.3e-06}^{***}$	0.336
	PWP-GT	-2.305	0.305	0.299	-2.809	-1.676	-4.266	$2e-05^{***}$	0.337
	Frailty	-3.677	0.361	88.29	-3.680	-2.466	1.000	$1.8e-11^{***}$	1.243

Signif. codes: '***' 0.001, '**' 0.01 '*' 0.05, '.' 0.1, ' ' 1
'Frailty +' modelos con término de fragilidad significativo

Al hallar las desviaciones estándar de los coeficientes vemos que en todos los casos obtenemos valores significativamente mayores que en el caso anterior. Esto se puede explicar pues al tomar la covariable dependiente del tiempo se le dificulta a los modelos aplicados estimar los coeficientes, lo que ocasiona una mayor variabilidad en los resultados.

BD3 y BD4

Para el tercer y cuarto grupo de bases de datos resultaron significativos todos los modelos ajustados excepto solo uno y en ambos casos los que obtuvieron los resultados más precisos fueron los modelos de fragilidad. Lo anterior era lo esperado teniendo en cuenta que en la simulación de estos datos se consideró que para individuos distintos existía un exceso de riesgo de presentar el evento el cual no se explicaba con la covariable observada. Los modelos de fragilidad se enfocan específicamente en tratar esta susceptibilidad heterogénea.

Para el conjunto de bases de datos 3 la media de los coeficientes estimados por el modelo de fragilidad fue -1.034 en el caso de 20 individuos, -1.001 para 30 individuos y para 50 individuos

se obtuvo -1.027, siendo el valor real -1 en los 3 casos considerados. En la tabla siguiente se muestran los resultados obtenidos y las desviaciones estándar de las estimaciones.

Tabla 3.7: Media de las estimaciones de los modelos para las Bases de datos 3

N	Modelo	coef(X)	se(coef(X))	rob.se	L.95	U.95	Est.	p-valor	sd(coef)
n=20									
	AG	-1.037	0.131	0.164	-1.316	-0.685	-2.624	0.009^{**}	0.249
	PWP-CP	-0.900	0.269	0.218	-1.247	-0.428	-1.446	0.148	0.364
	PWP-GT	-1.048	0.209	0.220	-1.405	-0.565	-1.776	0.076	0.345
	Frailty+	**-1.034**	**0.141**	**38.51**	**-1.377**	**-0.616**	**1.000**	$\mathbf{0.043}^{*}$	**0.241**
n=30									
	AG	-0.981	0.103	0.146	-1.237	-0.669	-2.696	0.007^{**}	0.211
	PWP-CP	-0.772	0.194	0.168	-1.063	-0.422	-1.845	0.065	0.227
	PWP-GT	-0.978	0.158	0.193	-1.299	-0.562	-2.183	0.029^{*}	0.292
	Frailty+	**-1.001**	**0.112**	**53.09**	**-1.299**	**-0.655**	**1.000**	$\mathbf{0.031}^{*}$	**0.197**
n=50									
	AG	-1.028	0.075	0.114	-1.231	-0.790	-4.014	$5.9e-05^{***}$	0.174
	PWP-CP	-0.759	0.140	0.130	-0.993	-0.495	-3.276	0.001^{*}	0.162
	PWP-GT	-1.070	0.116	0.158	-1.285	-0.680	-3.411	0.0007^{***}	0.213
	Frailty+	**-1.027**	**0.087**	**7.803**	**-1.256**	**-0.773**	**1.000**	$\mathbf{1.28e-06}^{***}$	**0.150**

Signif. codes: '***' 0.001, '**' 0.01 '*' 0.05, '.' 0.1, ' ' 1

'Frailty +' modelos con término de fragilidad significativo

Por su parte, para el conjunto de bases de datos 4 los coeficientes reales fueron, para el caso de 20 individuos -3, para 30 individuos -4 y para 50 se tomó igual a -5, obteniéndose como estimaciones medias los valores -3.042, -3.720 y -4.468 respectivamente. A continuación tenemos la tabla que resume para cada uno de los casos considerados los modelos aplicados, sus ajustes, los resultados obtenidos y las desviaciones estándar de los coeficientes estimados.

Tabla 3.8: Media de las estimaciones de los modelos para las Bases de datos 4

N	Modelo	coef(X)	se(coef(X))	rob.se	L.95	U.95	Est.	p-valor	sd(coef)
n=20									
	AG	-2.506	0.714	0.784	-3.401	-0.844	-1.668	$0.095^{\cdot}$	0.944
	PWP-CP	-2.524	0.864	0.726	-3.261	-0.916	-1.675	$0.094^{\cdot}$	1.023
	PWP-GT	-2.692	0.848	0.724	-3.438	-1.083	-1.957	0.050^{*}	0.997
	Frailty+	**-3.042**	**0.782**	**13.82**	**-3.472**	**-1.058**	**1.000**	$\mathbf{0.043}^{*}$	1.963
n=30									
	AG	-3.486	0.786	0.873	-4.557	-1.650	-2.688	0.007^{**}	0.903
	PWP-CP	-3.326	0.893	0.749	-4.187	-1.681	-2.853	0.004^{**}	0.935
	PWP-GT	-3.410	0.884	0.769	-4.314	-1.734	-2.244	0.025^{*}	0.888
	Frailty+	**-3.720**	**0.831**	**18.71**	**-4.408**	**-1.700**	**1.000**	$\mathbf{0.003}^{**}$	1.640
n=50									
	AG	-4.289	0.776	0.942	-5.160	-2.207	-3.204	0.001^{***}	1.185
	PWP-CP	-3.860	0.832	0.734	-4.645	-2.198	-3.862	0.0001^{***}	1.023
	PWP-GT	-3.921	0.823	0.736	-4.753	-2.273	-3.672	0.0002^{***}	0.977
	Frailty+	**-4.468**	**0.826**	**26.68**	**-4.911**	**-2.289**	**1.000**	$\mathbf{0.0002}^{***}$	1.783

Signif. codes: '***' 0.001, '**' 0.01 '*' 0.05, '.' 0.1, ' ' 1

'Frailty +' modelos con término de fragilidad significativo

Notemos que en este último caso las estimaciones no son tan exactas como las de los conjuntos de la base de datos anterior a pesar de que sí se mantiene que las mejores estimaciones se obtienen con los modelos que consideran fragilidad como era lo esperado. También debemos notar que, al igual que sucedió con las desviaciones estándar para las estimaciones de los conjuntos de las bases de datos 2, en este caso estos valores son mucho mayores que los encontrados en el caso anterior. Estas situaciones evidencian que cuando la covariable se toma dependiente del tiempo, el ajuste se complejiza y las estimaciones pierden exactitud.

3.2. Datos reales

Cáncer es el nombre que se da a un conjunto de enfermedades genéticas relacionadas, las cuales son causadas por cambios en los genes que controlan la forma en que funcionan nuestras células, especialmente la forma en que crecen y se dividen. Todos los tipos de cáncer se originan cuando algunas de las células del cuerpo empiezan a dividirse sin control, sobrepasando así en número a las células normales. Estas células cancerígenas se diseminan a los tejidos cercanos, lo que hace que al cuerpo le resulte difícil funcionar de la manera que debería hacerlo. Existen muchos tipos de cáncer los cuales se clasifican dependiendo del lugar del organismo donde se haya originado, puede ser por ejemplo en los pulmones, en el seno o hasta en la sangre. Estos diferentes tipos de cáncer tienen algunas similitudes, no obstante, son diferentes en la manera en que crecen y se propagan.

El cáncer colorrectal se produce cuando tumores se forman en el revestimiento del colon o del recto, los cuales forman parte del intestino grueso. Este es común tanto en hombres como en mujeres y el riesgo de desarrollarlo aumenta considerablemente después de los 50 años. En la mayoría de los casos, el diagnóstico del cáncer localizado es por colonoscopia. Muchos factores relacionados con el estilo de vida, tales como la obesidad, la inactividad física, el tabaquismo y el alcoholismo, han sido vinculados al cáncer colorrectal, identificando estos como factores que pueden aumentar las probabilidades de llegar a padecer esta enfermedad. Existen también factores de riesgo sobre los cuales no podemos influir como lo son antecedentes personales de cáncer colorrectal, de pólipos colorrectales o de enfermedades inflamatorias del intestino. Actualmente no hay una forma para ciertamente prevenir el cáncer colorrectal, sin embargo se pueden tomar medidas que generalmente ayudan a reducir el riesgo de padecerlo. Entre estas medidas se encuentran realizarse periódicamente pruebas de detección del cáncer colorrectal, cuidar su peso corporal, realizar actividades físicas regularmente y mantener una buena alimentación. Otras medidas pueden ser tomar vitaminas, calcio y magnesio y cuando utilicemos medicamentos antiinflamatorios comprobar que estos no sean esteroideos. El tratamiento es por lo general quirúrgico, y en muchos casos es seguido por quimioterapia.

3.2.1. Problemática y datos

Debido a la importancia de conocer todo lo posible sobre esta enfermedad, las tendencias de incidencia, mortalidad y sobrevivencia, así como sus factores de riesgo se han estudiado ampliamente. Se han analizado también las diferencias socioeconómicas y de género en la incidencia y

la mortalidad por este tipo específico de cáncer. Uno de estos estudios realizados con el objetivo de evaluar los efectos del género y la quimioterapia en el reingreso hospitalario después de la cirugía fue llevado a cabo entre pacientes diagnosticados de cáncer colorrectal que asistían al hospital universitario Bellvitge de Barcelona [6]. En este estudio se analizaron un total de 403 pacientes los cuales ya habían sido intervenidos en el periodo comprendido entre enero de 1996 y diciembre de 1998 y los cuales presentaron un nuevo diagnóstico de cáncer colorrectal. A estos se les realizó un seguimiento hasta junio del 2002 y se les controlaron varias variables las cuales se detallan a continuación.

Variables medidas en el estudio:

id: Identificador de cada sujeto. (Se encuentra repetido para cada recurrencia)

enum: Número de la readmisión.

t.start: Inicio del intervalo. (Toma el valor 0 o el tiempo de recurrencia anterior)

t.stop: Tiempo de censura o tiempo en el que ocurre el evento nuevamente.

time: Tiempo de internación o tiempo de censura final.

event: Situación de rehospitalización por evento. (Todos los eventos son 1 para cada sujeto excepto el último que es 0)

chemo: Variable que indica si el paciente recibió o no quimioterapia. (Valores: 1: No; 2: Si)

sex: Género. (Valores: 1: Masculino; 2: Femenino)

dukes: Estado tumoral de Dukes. (Valores: 1: A-B; 2: C; 3: D)

charlson: Índice de Comorbilidad de Charlson. Covariable dependiente del tiempo. (Valores: 0: índice 0; 1: índice 1-2; 3: índice ≥ 3)

death: Indicador de sobrevivencia. (Valores: 1: Muerto; 0: Vivo)

El conjunto de variables anteriores y sus mediciones para cada individuo se encuentran disponibles con el comando *readmission* de la librería *frailtypack* [24] en `R`.

Nuestro objetivo es utilizar los datos recogidos por estas variables para realizar modelos de sobrevivencia y modelos de eventos recurrentes con el fin de analizar el evento definido por la readmisión de cada paciente en el hospital y finalmente interpretar los resultados obtenidos.

3.2.2. Resultados

En las próximas subsecciones mostraremos los resultados de nuestro análisis, el cual se realizó integramente en `R`. Para facilitar la comprensión de estos no se incluyen los códigos.

Generalidades

Despues de cargar las librerías que utilizaremos y los datos, modificamos las variables `death` y `evento`, volviéndolas variables categóricas. Describimos en las siguientes dos tablas las características de cada una de las 11 variables que componen nuestra base de datos.

Tabla 3.9: Variables numéricas

Ident	Min	1st Qu.	Median	Mean	3er Qu.	Max
`enum`	1	1	2	2.641	3	23
`t.start`	0	0	25	254.5	390	2175
`t.stop`	1	206	607	734.5	1229	2176
`time`	1	21	216	480	880	2175

Tabla 3.10: Variables categóricas

Ident	Valores posibles (Observaciones)		
`event`	1 (458)	0 (403)	
`chemo`	Non Treated (468)	Treated (393)	
`sex`	Male (549)	Female (312)	
`dukes`	A-B (324)	C (331)	D (206)
`charlson`	0 (577)	1-2 (46)	3 (238)
`death`	0 (749)	1 (112)	

Se cuenta en total con 861 observaciones para cada una de las 11 variables, las cuales corresponden a un total de 403 individuos diferentes. Un aspecto importante a destacar es que no contamos con ningún valor `NA` en nuestra base de datos. A continuación mostramos una previsualización de esta en la cual hemos abreviado `Non Treated`, `Treated`, `Female` y `Male` por `NT`, `T`, `F` y `M` respectivamente.

	`id`	`enum`	`t.start`	`t.stop`	`time`	`event`	`chemo`	`sex`	`dukes`	`charlson`	`death`
1	1	1	0	24	24	1	`T`	`F`	`D`	3	0
2	1	2	24	457	433	1	`T`	`F`	`D`	0	0
3	1	3	457	1037	580	0	`T`	`F`	`D`	0	0
4	2	1	0	489	489	1	`NT`	`M`	`C`	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Realizamos también una gráfica con los eventos recurrentes por individuos en el tiempo en la cual también señalamos el tiempo de muerte. Se obtuvo lo mostrado en la Figura 3.5.

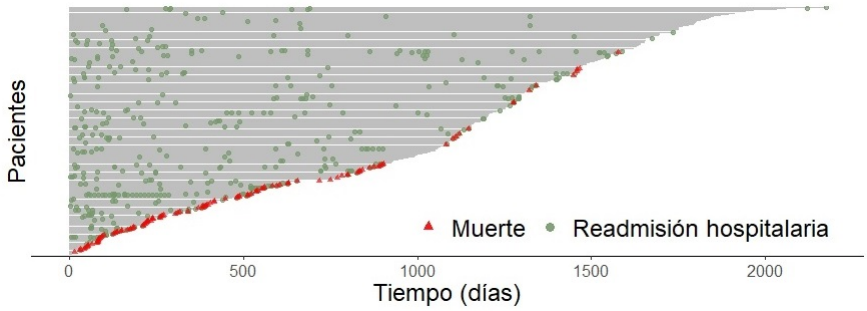

Figura 3.5: Eventos recurrentes por pacientes

Por último analizamos la cantidad de pacientes con el mismo número de readmisiones y hallamos su tiempo de sobrevivencia promedio, obteniendo los siguientes datos mostrados en la Tabla 3.11 y en la Figura 3.6.

Tabla 3.11: Cantidad de readmisiones y número de casos

Readm.	1	2	3	4	5	6	7	9	10	11	12	18	23
Casos	181	114	45	27	15	10	4	1	2	1	1	1	1

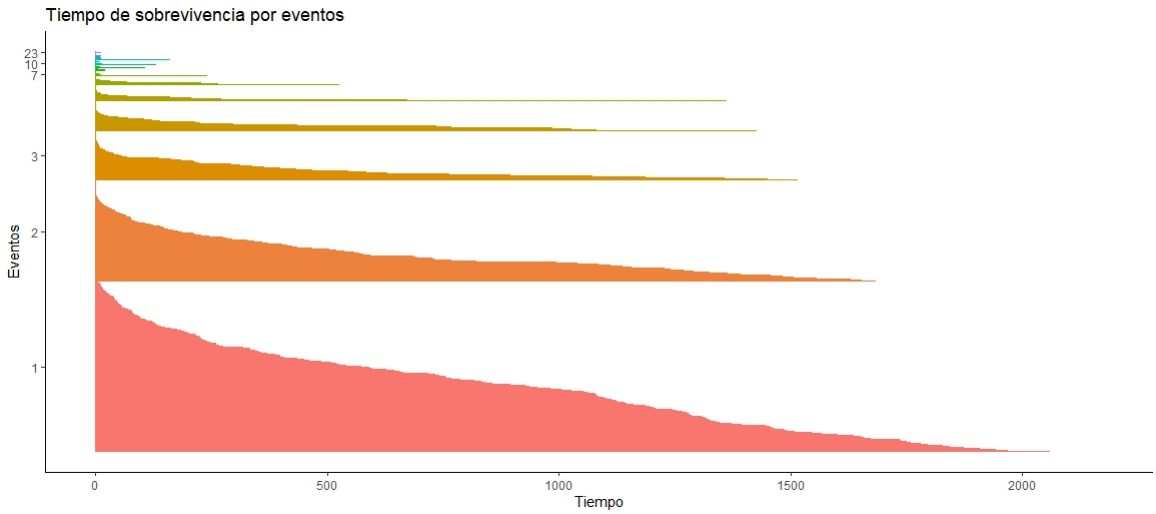

Figura 3.6: Cantidad de readmisiones y número de casos

Métodos de análisis de sobrevivencia

En esta subsección mostraremos los resultados obtenidos al estimar la función de sobrevivencia y la función de riesgo. Lo anterior lo desarrollaremos desde tres puntos diferentes, utilizando técnicas descriptivas no paramétricas, empleando algunos métodos paramétricos y finalmente utilizando la regresión de riesgos proporcionales de Cox, clasificada como un método semiparamétrico.

Técnicas descriptivas no paramétricas

Se realizó la estimación de la función de sobrevivencia (empleando el método de Kaplan-Meier) y de la función de riesgo acumulado utilizando todos los datos y considerándolos independientes. Se obtuvieron tres funciones de cada tipo, ya que el análisis lo realizamos considerando toda la muestra y tomando la muestra dividida por sexo. A continuación graficamos las funciones obtenidas en cada caso con sus respectivos intervalos de confianza.

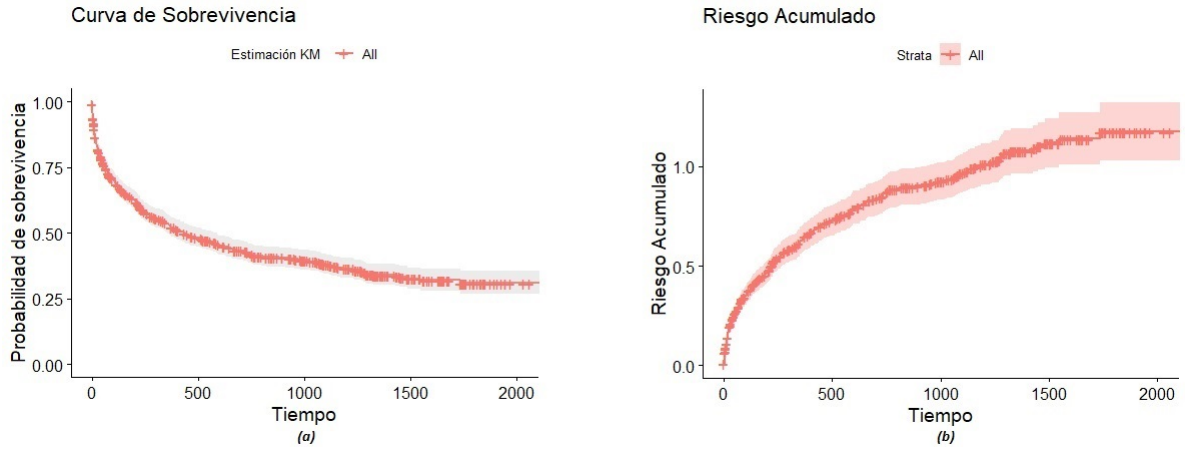

Figura 3.7: Curva de sobrevivencia (a) y función del riesgo acumulado (b) para el análisis de la muestra completa

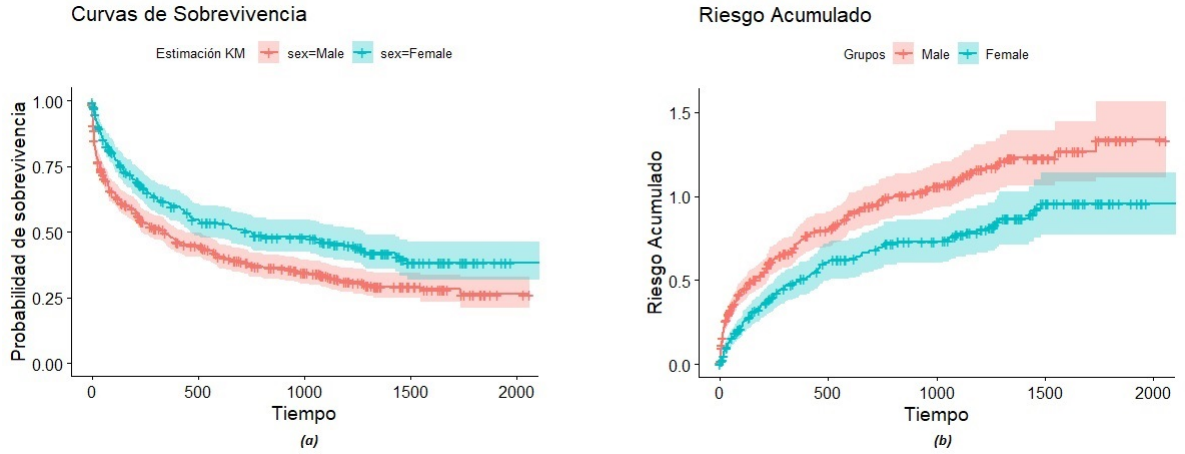

Figura 3.8: Curva de sobrevivencia (a) y función del riesgo acumulado (b) para el análisis de la muestra separada por sexo

Debido a que tenemos dos grupos, pacientes femeninos y masculinos, realizamos la siguiente prueba de hipótesis:

$$H_0 : S_M(t) = S_F(t) \qquad H_1 : S_M(t) \neq S_F(t) \qquad \forall t$$

La anterior se realizó con el objetivo de verificar si existen diferencias significativas entre los comportamientos de las funciones de sobrevivencia de cada grupo. Para la prueba LogRank realizada se obtuvo un p-valor aproximado de $8 * 10^{-5}$, por lo que rechazamos la hipótesis de

igualdad de las funciones de sobrevivencia concluyendo que estas tienen comportamientos diferentes en los dos grupos analizados.

Métodos paramétricos

Los métodos paramétricos son aquellos que ajustan un modelo con una distribución conocida. Para nuestros datos consideramos el ajuste de modelos con siete distribuciones conocidas, ellas fueron la Distribución Exponencial, Weibull, Log-Logística, Gompertz, F generalizada, Log-Normal y Gamma. Nuevamente realizamos tres análisis, uno para cada sexo y otro que incluyera a los datos de ambos sexos. A continuación mostramos las gráficas de las funciones de sobrevivencia y de riesgo acumulado en cada caso, así como tablas con las medidas de ajuste de cada modelo ajustado.

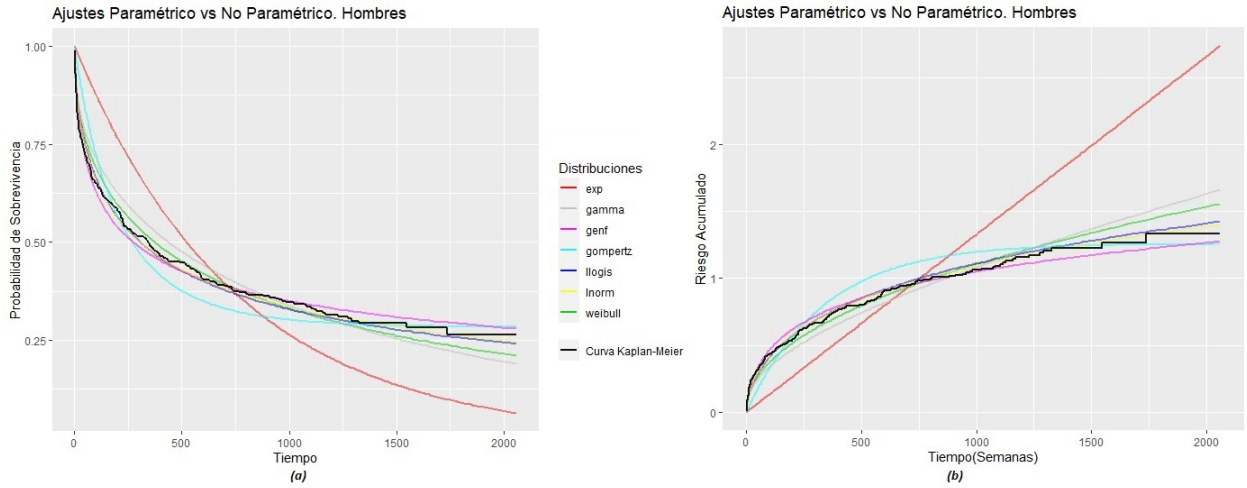

Figura 3.9: Curvas de sobrevivencia (a) y funciones del riesgo acumulado (b) halladas con los métodos paramétricos vs la Curva de Kaplan-Meier. Hombres

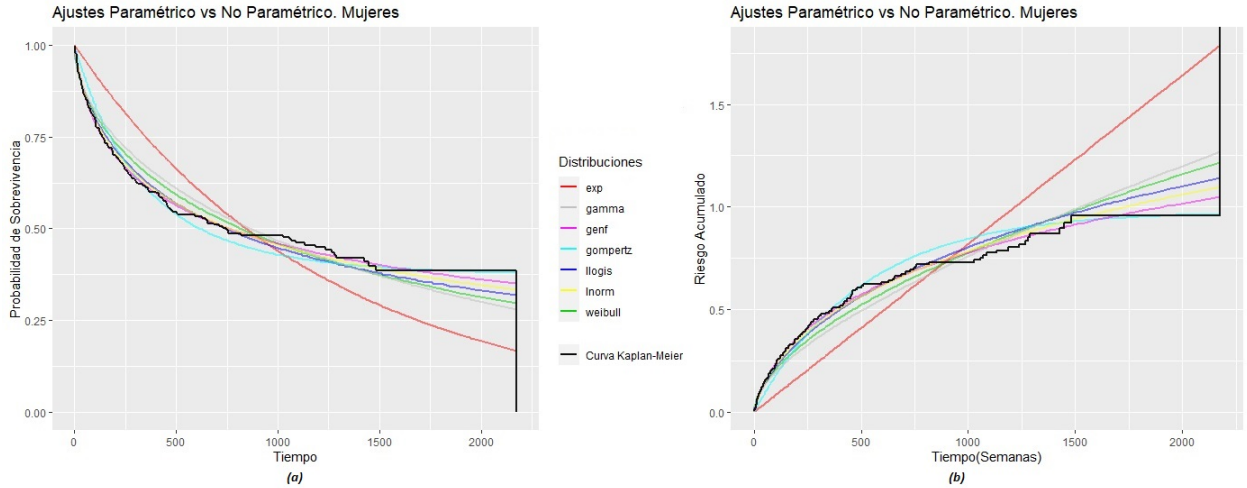

Figura 3.10: Curvas de sobrevivencia (a) y funciones del riesgo acumulado (b) halladas con los métodos paramétricos vs la Curva de Kaplan-Meier. Mujeres

Tabla 3.12: Medidas de ajuste para los modelos paramétricos. Hombres

Medida	exp	weibull	llogis	gompertz	genf	lnorm	gamma
AIC	4728.025	4392.678	4374.029	4496.783	4352.422	4356.888	4414.460
BIC	4732.333	4401.294	4382.645	4505.399	4369.654	4365.504	4423.076
LogLik	-2363.012	-2194.339	-2185.015	-2246.391	-2172.211	-2176.444	-2205.230

Tabla 3.13: Medidas de ajuste para los modelos paramétricos. Mujeres

Medida	exp	weibull	llogis	gompertz	genf	lnorm	gamma
AIC	2400.961	2325.234	2316.677	2329.929	2313.829	2310.934	2332.116
BIC	2404.704	2332.720	2324.163	2337.415	2328.801	2318.420	2339.602
LogLik	-1199.481	-1160.617	-1156.338	-1162.965	-1152.915	-1153.467	-1164.05

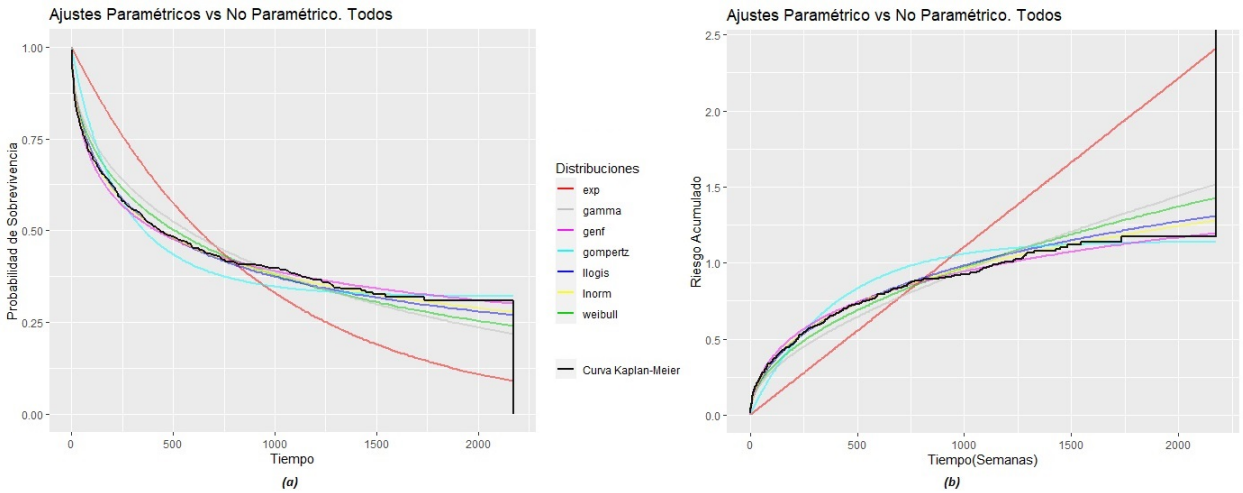

Figura 3.11: Curvas de sobrevivencia (a) y funciones del riesgo acumulado (b) halladas con los métodos paramétricos vs la Curva de Kaplan-Meier. Todos

Tabla 3.14: Medidas de ajuste para los modelos paramétricos. Todos

Medida	exp	weibull	llogis	gompertz	genf	lnorm	gamma
AIC	7151.415	6735.371	6711.091	6846.234	6688.536	6690.384	6762.414
BIC	7156.173	6744.887	6720.608	6855.750	6707.569	6699.901	6771.930
LogLik	-3574.707	-3365.686	-3353.546	-3421.117	-3340.268	-3343.192	-3379.207

Después de analizar los resultados obtenidos, podemos concluir que los modelos paramétricos que mejor se ajustan a nuestros datos desde un enfoque global son los que utilizan la distribución Log-Normal. Se presenta en la Tabla 3.19 los coeficientes estimados y sus respectivos intervalos de confianza para cada modelo ajustado.

Al incluir covariables en los modelos se afecta la convergencia de los mismos, y en los casos en que sí se llega a la convergencia los coeficientes estimados no varían considerablemente, por lo que se decidió no incluir covariables en estos modelos.

Tabla 3.15: Estimaciones de cada modelo y características

		Estimaciones				Observaciones		
		est	L95%	U95%	se	N	Eventos	Cens.
Hombres	meanlog	5.722	5.449	5.994	0.139			
	sdlog	2.779	2.558	3.019	0.117			
						549	310	239
Mujeres	meanlog	6.635	6.293	6.977	0.175			
	sdlog	2.437	2.156	2.755	0.152			
						312	148	164
Todos	meanlog	6.0792	5.8601	6.2982	0.1118			
	sdlog	2.7308	2.5497	2.9248	0.0956			
						861	458	403

Métodos semiparamétricos

Para concluir con los métodos de análisis de sobrevivencia ajustaremos modelos de regresión de Cox, los cuales son el método semiparamétrico más empleado en estos tipos de análisis.

En el análisis realizado incluimos las covariables dependientes del tiempo, el tiempo en que ocurre la recaída y el tiempo final que se tiene para cada uno. También ajustamos modelos incluyendo todas las combinaciones posibles de covariables, sin embargo, la mayoría de las covariables y sus interacciones resultaron no significativas. En los análisis realizados solo resultaron significativas los efectos principales de las variables `sex` y `dukes` [1], por lo que decidimos quedarnos con un modelo que solo incluyera a estas. A continuación presentamos los coeficientes estimados del modelo, sus errores estándar, intervalos de confianza y puntuaciones `z`.

Tabla 3.16: Estimaciones del modelo seleccionado

	coef	exp(coef)	se(coef)	rob. se	L .95	U .95	z	Pr($>\|z\|$)
sexF	-0.4848	0.6158	0.1008	0.1685	0.4426	0.8568	-2.878	0.00401^{**}
dukesC	0.5022	1.6524	0.1120	0.1829	1.1546	2.3649	2.746	0.00604^{**}
dukesD	1.6188	5.0472	0.1246	0.2369	3.1725	8.0296	6.833	$8.29e-12^{***}$
	Signif. codes: '***' 0.001, '**' 0.01 '*' 0.05, '.' 0.1, ' ' 1							

Una vez seleccionado el modelo realizamos la validación del supuesto de riesgos proporcionales, donde, al realizar una prueba de hipótesis obtuvimos un p-valor global mayor a 0.05,

[1] Este resultado coincide con resultados de análisis realizados por otros autores. Ver [6]

por lo que no existe evidencia para suponer que este supuesto no se cumple. De igual manera comprobamos que se cumplieran los supuestos relacionados con la linealidad y los residuales.

Para concluir graficamos las curvas de sobrevivencia estimadas por el modelo seleccionado.

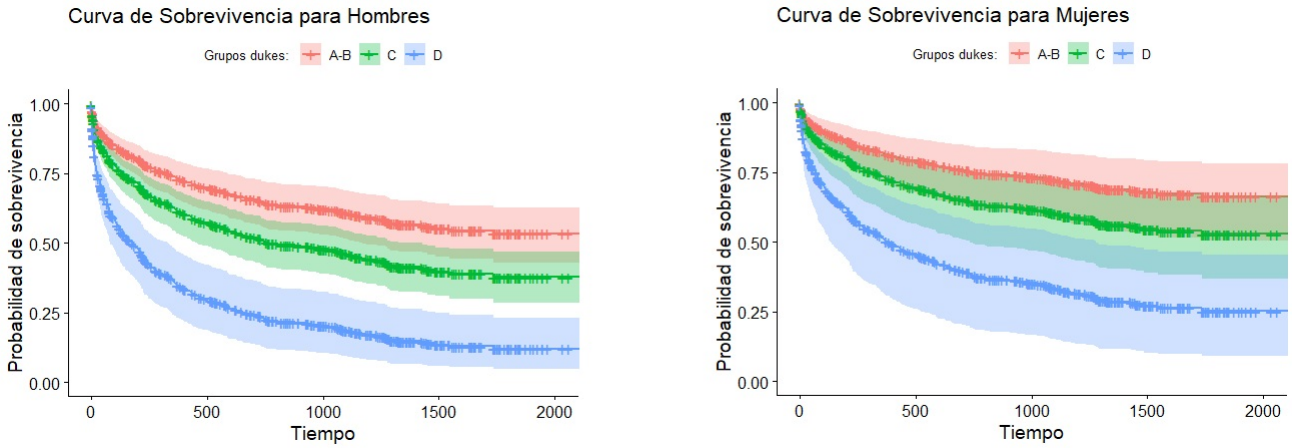

Figura 3.12: Curvas de sobrevivencia para hombres y mujeres estimadas mediante el modelo de regresión de Cox

Métodos de análisis de eventos recurrentes

Basándonos en los resultados anteriores, trabajaremos los modelos de eventos recurrentes solamente con las covariables `sex`, `dukes`, aquellas que señalan el tiempo en que ocurren los sucesos y la que indica el estado del individuo en cada uno de estos tiempos.

Modelo de Andersen-Gil

Al ajustar un modelo de Andersen-Gil con las covariables seleccionadas, obtenemos los siguientes resultados:

Tabla 3.17: Estimaciones del modelo de AG

	coef	exp(coef)	se(coef)	rob. se	L .95	U .95	z	Pr($>\|z\|$)
sex F	-0.4842	0.6162	0.1008	0.1682	0.4431	0.8569	-2.878	0.00400**
dukesC	0.5018	1.6518	0.1120	0.1828	1.1545	2.3633	2.746	0.00604**
dukesD	1.6172	5.0390	0.1246	0.2365	3.1699	8.0102	6.838	$8.01e-12$***
	Signif. codes: '***' 0.001, '**' 0.01 '*' 0.05, '.' 0.1, ' ' 1							

Notemos que todos los coeficientes resultan significativos y que como se espera en los modelos marginales, los errores robustos estándar son mayores que los errores estándar del modelo base. Al comparar este modelo con el modelo de riesgos proporcionales de Cox ajustado anteriormente podemos comprobar que los resultados obtenidos con ambos modelos son muy

similares.

Una dificultad que presentan estos modelos es que se ignora la variable `enum`, la cual identifica el número de la recaída del paciente. Debido a lo anterior este modelo asume que todos los eventos son iguales, lo cual resulta un supuesto muy fuerte en la mayoría de las ocasiones. En los datos con los que estamos trabajando específicamente, este supuesto implicaría que el riesgo de una nueva hospitalización es el mismo para los individuos, sin importar si han estado hospitalizados antes y si es así, cuántas veces lo han estado. Es de esperar en este estudio que el riesgo de una nueva hospitalización dependa de los eventos anteriores, por lo que este modelo no sería la mejor opción para analizar los datos.

Modelos de Prentice, Williams y Peterson

Los modelos de Prentice, Williams y Peterson resuelven la dificultad que planteábamos anteriormente, ya que estos están preparados para incorporar el uso de estratos dependientes del tiempo, lo que significa que la función de riesgo subyacente puede variar de un evento a otro.

Utilizando los datos disponibles ajustamos los dos modelos de PWP, uno basado en procesos de conteo (CP) y el otro en intervalos de tiempo (GT), de los cuales presentamos los resultados en las siguientes tablas.

Tabla 3.18: Estimaciones del modelo de PWP-CP

	coef	exp(coef)	se(coef)	rob. se	L .95	U .95	z	Pr($>\|z\|$)
sexF	-0.3223	0.7245	0.1072	0.1123	0.5813	0.9029	-2.869	0.00411^{**}
dukesC	0.3294	1.3901	0.1175	0.1319	1.0734	1.8004	2.497	0.01253^{*}
dukesD	1.0756	2.9318	0.1395	0.1424	2.2178	3.8756	7.553	$4.24e-14^{***}$
	Signif. codes: '***' 0.001, '**' 0.01 '*' 0.05, '.' 0.1, ' ' 1							

Tabla 3.19: Estimaciones del modelo de PWP-GT

	coef	exp(coef)	se(coef)	rob. se	L .95	U .95	z	Pr($>\|z\|$)
sexF	-0.34861	0.70567	0.10240	0.09385	0.5871	0.8482	-3.715	0.000204^{***}
dukesC	0.36386	1.43888	0.11363	0.12541	1.1253	1.8398	2.901	0.003715^{**}
dukesD	0.95327	2.59418	0.13387	0.13615	1.9866	3.3876	7.001	$2.53e-12^{***}$
	Signif. codes: '***' 0.001, '**' 0.01 '*' 0.05, '.' 0.1, ' ' 1							

Como podemos notar en ambos modelos todos los coeficientes son significativos, y, a pesar de que los resultados son diferentes, estas diferencias no son muy grandes. Vemos también que los intervalos de confianza se enciman y el sentido del efecto de las variables no cambia.

Debemos recordar que el modelo de PWP-CP evalúa el efecto de una covariable para el k-ésimo evento desde el momento de la entrada del paciente en el estudio; mientras que el modelo PWP-GT evalúa el efecto de una covariable para el k-ésimo evento desde el momento en que se presenta el evento anterior. Debido a lo anterior la selección del modelo adecuado dependerá de nuestras preguntas de investigación. Para estos datos, por ejemplo, la escala de procesos de conteo puede ser de interés si el proceso de la enfermedad del paciente se considera como un todo, mientras que los tiempos de intervalos será la escala recomendada cuando buscamos estudiar los episodios de la enfermedad.

Modelos de fragilidad

Finalmente ajustamos un modelo con un término de fragilidad para dar cuenta de la heterogeneidad entre los individuos del estudio. Al ajustar el modelo, obtuvimos los siguientes resultados:

Tabla 3.20: Estimaciones del modelo de fragilidad

	coef	se(coef)	se2	L.95	U.95	Chisq	DF	p
sexF	-0.3700	0.1128	0.1028	0.5538	0.8616	10.76	1.00	1.0e-03
dukesC	0.3202	0.1224	0.1129	1.0836	1.7509	6.84	1.00	8.9e-03
dukesD	1.1810	0.1445	0.1291	2.4544	4.3238	66.85	1.00	2.9e-16
frailty.(id)						59.22	43.92	6.1e-02

La idea de utilizar este modelo es que el efecto aleatorio describe un exceso de riesgo o fragilidad para individuos distintos, teniendo en cuenta la heterogeneidad no medida que no puede explicarse por las covariables observadas solamente. Este resulta útil cuando hay susceptibilidad heterogénea ante el riesgo de eventos recurrentes, situación que podemos encontrar en los datos con los que estamos trabajando.

Modelo final e interpretación

Basándonos en las características de los modelos ajustados y en las dificultades y ventajas de cada uno de ellos, decidimos que el que mejor se adapta a los datos presentados es el modelo de Prentice, Williams y Peterson utilizando el enfoque de procesos de conteo, ya que consideramos el proceso de la enfermedad como un todo.

Interpretando los coeficientes estimados, los cuales se muestran en la Tabla 3.18, obtenemos que:

- Ser mujer reduce en un 27 % el riesgo de tener una recaída y tener que ser hospitalizado.
- Aquellos pacientes que se encuentran clasificados en el estado tumoral de Dukes A o B, tienen probabilidades más bajas de presentar episodios recurrentes de recaídas que aquellos clasificados en el estado tumoral de Dukes C o D.

- Mientras más avanzado sea el estado tumoral según la clasificación de Dukes (teniendo en cuenta que el orden es A, B, C y D) aumenta la probabilidad de recaídas en pacientes, y con ello el número de eventos de cada uno.

Capítulo 4

Conclusiones

Después de realizado el siguiente trabajo podemos concluir que:

- Desde su surgimiento el análisis de sobrevivencia se consideró principalmente como un instrumento analítico para estudios biomédicos y demográficos. En los últimos 50 años y debido al desarrollo de técnicas informáticas más potentes, este tipo de análisis se ha expandido gradualmente a aréas tan variadas como la ingeniería, no obstante, su mayor aréa de aplicación sigue siendo el aréa de ciencias de la salud.

- La principal ventaja que tienen los métodos de análisis de sobrevivencia frente a los métodos estadísticos tradicionales es que estos sí permiten el tratamiento con datos que presenten censura y covariables que dependan del tiempo. Lo anterior explica por qué al analizar datos con estas características obtendremos resultados más fiables y exactos con estas técnicas.

- Un tipo particular de datos de sobrevivencia son aquellos obtenidos a partir de eventos recurrentes, o sea, eventos donde se considera que el suceso de interés puede ocurrir varias veces para un mismo individuo. El análisis de eventos recurrentes, desde un punto de vista estadístico, presenta dos desafíos importantes: la correlación intraindividual y, al igual que el análisis de sobrevivencia, el tratamiento de las covariables que varían en el tiempo.

- Se han propuesto varios enfoques para el análisis de datos de eventos recurrentes, los cuales, a diferencia de las técnicas estadísticas tradicionales e incluso de las técnicas de análisis de sobrevivencia típicas, permiten abordar el proceso que se estudia de manera más apropiada. A pesar de lo anterior, si no se tiene en cuenta la naturaleza y las característica de los datos en la elección del modelo a emplear, se pueden presentar deficiencias en los resultados como mal ajuste del modelo o interpretaciones que no tengan sentido.

- Se realizó un estudio creando bases de datos mediante simulaciones de eventos recurrentes con ciertas características específicas. A dichas bases de datos se les aplican los modelos estudiados con el objetivo de comparar el desempeño de cada uno y sus resultados obtenidos. Lo anterior nos permite corroborar la correspondencia entre cada modelo y ciertos rasgos propios de los eventos.

- Se concluye con un estudio realizado utilizando una base de datos real sobre pacientes con cáncer con el objetivo de realizar modelos de sobrevivencia y modelos de eventos recurrentes para estudiar el evento definido por la readmisión de cada paciente en el hospital. Al analizar los datos y sus características propias se concluye que el modelo que mejor se adapta en este caso era el modelo de Prentice, Williams y Peterson utilizando el enfoque de procesos de conteo (PWP-CP).

Bibliografía

[1] Altman, D. G. (1991). *Practical statistics for medical research.* Chapman & Hall, London and New York.

[2] Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100-1120.

[3] Barceló, M. A., & Saez, M. (2004). A Modification Of The EM Algorithm To Estimate An Andersen-Gill Gamma Frailty Model For Multivariate Failure Time Data. *Journal of Modern Applied Statistical Methods,* 3(2), 21.

[4] Bland, J. M., & Altman, D. G. (2004). The logrank test. *Bmj,* 328(7447), 1073.

[5] Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data. Comparisons and applications of some estimators and tests. *Journal of applied econometrics*, 1(1), 29-53.

[6] Charles-Nelson, A., Katsahian, S., & Schramm, C. (2019). How to analyze and interpret recurrent events data in the presence of a terminal event: An application on readmission after colorectal cancer surgery. *Statistics in medicine,* 38(18), 3476-3502.

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

[8] Efron, B., & Tibshirani, R. (1993) *An introduction to the bootstrap.* New York: Chapman & Hall/CRC.

[9] Glynn R. J., & Buring, J. E. (1996). Ways of measuring rates of recurrent events. *BJM,* 312, 364-367.

[10] Hardin, J., & Hilbe, J. (2007) *Generalized linear models and extensions.* 2nd ed. Texas: Stata Press.

[11] Hilbe. J. (2007). Negative Binomial Regression. *Cambridge University Press.*

[12] Hosmer, L., & Lemeshow, S. May (2008) *Applied Survival Analysis.* Wiley.

[13] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association,* 53(282), 457-481.

[14] Karim, M. R., & Islam, M. A. (2019). *Reliability and Survival Analysis.* Springer Singapore.

[15] Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text.* Springer.

[16] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics,* 34(1), 1-14.

[17] Lee, E. T., Desu, M. M., & Gehan, E. A. (1975). A Monte Carlo study of the power of some two-sample tests. *Biometrika,* 62(2), 425-432.

[18] Liu, X. (2012). *Survival analysis: models and applications.* John Wiley & Sons.

[19] McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* 2nd ed. New York: Chapman & Hall

[20] Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society:* Series A (General), 135(2), 185-198.

[21] Prentice, R. L., & Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics,* 861-867.

[22] Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2), 373-379.

[23] R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL **https://www.R-project.org/.**

[24] Rondeau, V., & Gonzalez, J. R. (2005). Frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer methods and programs in biomedicine,* 80(2), 154-164.

[25] Sy Han (Steven) Chiou and Chiung-Yu Huang (2021). *reReg: Recurrent Event Regression.* R package version 1.4.0. **https://CRAN.R-project.org/package=reReg**.

[26] Schmidt, P., & Witte, A. D. (2012). *Predicting recidivism using survival models.* Springer Science & Business Media.

[27] Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model.* Springer, New York, NY.

[28] Therneau, T. M., Grambsch, P. M., & Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of computational and graphical statistics,* 12(1), 156-175.

[29] Wang W, Fu H, Yan J (2021). *_reda: Recurrent Event Data Analysis_*. R package version 0.5.3, <URL: **https://github.com/wenjie2wang/reda**>.

[30] Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, 84(408), 1065-1073.

[31] Yadav, C. P., Sreenivas, V., Khan, M. A., & Pandey, R. M. (2018). An overview of statistical models for recurrent events analysis: a review. *Epidemiology (Sunnyvale)*, 8(4), 354.