VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

FACVLTAD DE CIENCIAS

GRADO EN ESTADÍSTICA

Trabajo de Fin de Grado

MODELOS MATEMÁTICOS VTILIZADOS EN ANÁLISIS DE SVPERVIVENCIA

MATHEMATICAL MODELS IN SVRVIVAL ANALYSIS

Autora: Andrea Sánchez Moreno

Tutora: María Jesús Rivas López

Salamanca, 2021

FACVLTAD DE CIENCIAS
GRADO EN ESTADÍSTICA

Trabajo de Fin de Grado

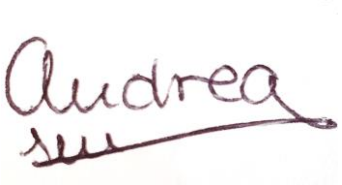
MODELOS MATEMÁTICOS VTILIZADOS EN ANÁLISIS DE SVPERVIVENCIA

MATHEMATICAL MODELS IN SVRVIVAL ANALYSIS

Autora: Andrea Sánchez Moreno

Tutora: María Jesús Rivas López

Andrea Sánchez Moreno



María Jesús Rivas López

Salamanca, 2021

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE GRÁFICOS

1. Introducción

Gran cantidad de estudios, realizados tanto en los últimos años como en las últimas décadas, se basan en el análisis del tiempo transcurrido hasta la ocurrencia de un suceso o evento de interés, ya sea el fallecimiento por cáncer, el cambio de un tratamiento a otro, o bien, el tiempo de eficacia de un fármaco. Este intervalo de tiempo se denomina tiempo de supervivencia o tiempo de evento.

Este tiempo podrá ser observado entre, un tiempo inicial, en el que da comienzo el estudio y se observan cada uno de los individuos que estén implicados en el mismo y, un tiempo o instante final, con el que se da por concluido el lapso de estudio.

En medicina, gran cantidad de enfermedades se someten a diferentes tipos de tratamientos según el grado en que se encuentre la misma, el nivel de mejoría de ésta, el sexo, la edad y muchas variables más, para obtener con ello, una mejor calidad de vida o aumentar el tiempo de vida del paciente. Como cada enfermedad puede depender de diferentes factores que la hagan más o menos perjudicial, los tratamientos o fármacos aplicados a los individuos favorecerán o no el progreso de la misma y, por lo tanto, el tiempo de supervivencia estimado variará para cada una de ellas. Estos procedimientos médicos, son utilizados para la realización de un estudio de análisis de supervivencia, llegando a pronosticar el tiempo de vida de cada uno de los pacientes.

Los sujetos se caracterizan por no tener un mismo tiempo inicial en el estudio, por ejemplo, un individuo puede comenzar el estudio en el tiempo inicial del mismo, es decir, en el tiempo 0, mientras que el tiempo de inclusión de otro individuo puede ser a los 5 años, meses o días del comienzo de estudio. Esto quiere decir que, aquel sujeto que inicia más tarde el estudio tendrá un menor periodo de observación y, por tanto, una menor probabilidad de que le ocurra el suceso dentro del periodo de estudio; pese a esta diferencia, ambos sujetos son analizados del mismo modo, ya que lo primordial es el tiempo transcurrido entre dicho tiempo inicial y el suceso. Este evento se produce una única vez y es el que hará concluir el tiempo de supervivencia del sujeto.

El objetivo es analizar el tiempo de permanencia en estudio de los individuos a los que le ocurre el evento, teniendo en cuenta que, en ocasiones, ciertos individuos lo abandonarán, otros superarán el tiempo de estudio, o les sucederá un evento diferente al propuesto; estos sujetos se denominan censuras. Por lo tanto, con los individuos que desarrollan el suceso predeterminado y parte de la información que proporcionan los individuos censurados, se pretende calcular la probabilidad de sobrevivir durante un tiempo concreto, es decir, obtener la probabilidad de que a los sujetos les ocurra el evento de interés antes de un tiempo determinado.

Actualmente, el análisis de supervivencia es fundamental en muchas aplicaciones, tanto en Medicina, como en Ingeniería, Industria, Economía, Marketing, … multitud de disciplinas en las que el objetivo de todas ellas es proporcionar un tiempo de supervivencia estimado para el objeto de estudio, el cual tiene unas características determinadas.

[1]

Una de las razones más importantes por las que se aplica el análisis de supervivencia, en muchas ocasiones, es porque la variable respuesta o variable de tiempo de supervivencia se caracteriza por la presencia de asimetría, por lo que, esta variable no puede seguir el modelo simétrico de una distribución normal. En la mayoría de las situaciones, la distribución será asimétrica hacia la derecha o asimétrica positiva, ya que, cuanto mayor es el tiempo de permanencia en el estudio, mayor será la probabilidad de supervivencia del conjunto de individuos.

Otro motivo por el que tiene tanta importancia este tipo de análisis es la capacidad de estudiar las denominadas censuras, anteriormente mencionadas, tiempos de evento de sujetos que no se observan durante todo el periodo de estudio; se puede conocer el tiempo de incorporación al estudio pero se desconoce el tiempo de ocurrencia del evento. Su tiempo de censura está determinado por un suceso aleatorio, distinto al fijado por el investigador; bien porque le ocurra un evento que le saque del estudio sin ser el evento que estamos observando, y por tanto, su tiempo de censura es ese tiempo de abandono del estudio, o bien, porque el evento le ocurre tras la finalización del estudio, y por tanto, se le otorga como tiempo de censura el tiempo final del estudio.

Cabe destacar en análisis de supervivencia los denominados modelos multi-estado y multi-evento. El primero de ellos se considera cuando existen diferentes estadios posibles para un evento, es decir, cuando se tienen diferentes fases en los que dividir el grado de cáncer, el individuo se mantiene en el estudio, a no ser que se dé el evento, ya que puede pasar de la fase III del grado de cáncer a la fase II y permanecer vivo. Por otro lado, el modelo multi-evento se tiene cuando existe más de un evento posible con el que dar por finalizado el tiempo de estudio de un paciente; en este caso, el evento observado no es dicotómico y se tendrán *k* posibles estadios finales. Pese a ser una parte más del análisis de supervivencia, no se considerarán más allá durante este trabajo.

Para describir mediante modelos matemáticos la variable respuesta o "tiempo hasta un evento", se utilizarán las variables explicativas o covariables, suponiendo siempre no dependientes del tiempo, para comprobar su influencia en el tiempo de supervivencia de los diferentes individuos del estudio.

En cuanto a la aplicación práctica de los métodos que se van a describir durante todo el trabajo, se realizará a través del programa estadístico R, que cuenta con numerosas funciones con las que llevar a cabo el análisis de supervivencia mediante paquetes como *"survival"*, *"KMsurv"*, *"survMisc"* o *"survminer"*. Las bases de datos utilizadas para los análisis son *"btrial"* y *"bfeed"*, las cuales vienen implementadas en el propio programa, en concreto en el paquete estadístico *"KMsurv"*.

El principal objetivo de este trabajo es la búsqueda de los modelos matemáticos más destacados en el estudio de la variable "tiempo hasta un evento", con el fin de conocer la influencia de ciertas covariables en el tiempo de supervivencia, siendo estas no dependientes del tiempo. Además, también analizar mediante procedimientos prácticos, los modelos destacados para análisis de supervivencia.

El presente trabajo se dividirá en seis secciones, comenzando con esta pequeña introducción hacia un punto de vista general del análisis de supervivencia, posteriormente, en el segundo capítulo, se definen los conceptos básicos de la supervivencia, haciendo inferencia en las censuras, las funciones claves, como lo son las funciones de densidad, de distribución, de supervivencia, de riesgo y de riesgo acumulado, y por último las distribuciones más destacadas (Exponencial y Weibull), formando la parte paramétrica del proyecto. Continuando con el desarrollo teórico, pasamos al análisis paramétrico, formado por el método de Kaplan-Meier y las pruebas utilizadas para una comparación entre curvas de supervivencia, en las que cabe destacar el test Log-rank, el test de Wilcoxon y el test de Tarone-Ware. Finalizando con los apartados teóricos, se tiene la regresión de Cox, análisis semiparamétrico con el que se crea un modelo formado por una componente paramétrica y otra no paramétrica. Para la sección número 5, se aplica mediante el software R, un análisis práctico de todos los métodos teóricos desarrollados previamente y el trabajo se completa con unas conclusiones finales del mismo.

2. Conceptos básicos de análisis de supervivencia

Para comenzar a estudiar el tema en profundidad, previamente se van a desarrollar unos conceptos que facilitan la compresión del análisis de supervivencia.

El siguiente esquema resume muy gráficamente lo que es el análisis de supervivencia. Por un lado, la duración del estudio, desde una fecha inicial hasta una fecha final. Dentro de ese tiempo de estudio, se tienen los tiempos de inclusión de cada individuo, en la mayoría de las ocasiones, dispares entre sí, dado que las entradas suelen ser de manera escalonada; y además, las fechas de la última observación, que pueden ser por causa del evento o por caso de censura, donde el sujeto se pierde o abandona el estudio o llega a su fin sin haberle sucedido el evento.

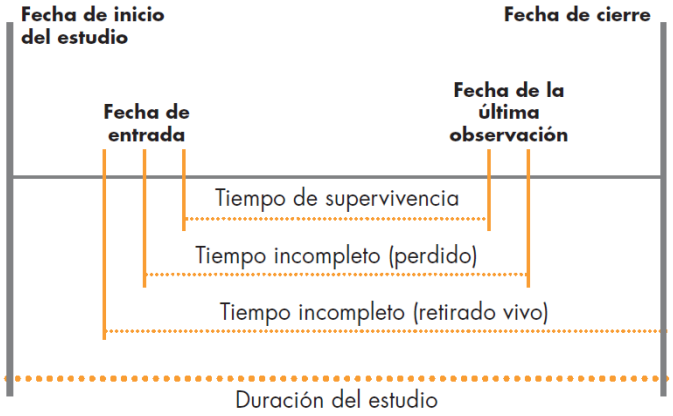


Figura 1. Esquema de un análisis de supervivencia.

El análisis de supervivencia consiste en determinar el efecto que produce una variable independiente cuando la variable dependiente está definida como el tiempo hasta el suceso de un evento.

Durante el proceso de análisis se tienen tres tipos de variables implicadas, por un lado, la variable respuesta, que viene definida por el tiempo en el que se mantienen los participantes en estudio; por otro lado, la variable de censura, aquella que determina la ocurrencia del evento de estudio o no; y, por último la variable explicativa o variable independiente, con la que se estudiará la influencia de diferentes factores sobre la variable respuesta. El conjunto de los dos primeros tipos de variables, formarán la variable dependiente, tiempo hasta que ocurre un evento o suceso.

El tiempo hasta el evento de interés es la variable relevante en análisis de supervivencia. Ésta representa el tiempo de seguimiento de cada individuo y debe estar perfectamente definida antes de comenzar el estudio, además de definir su momento inicial. Por lo que, se define el tiempo de observación de un individuo como el tiempo transcurrido desde que un individuo entra en estudio hasta el suceso del evento, o bien, hasta el fin del tiempo de estudio o de abandono.

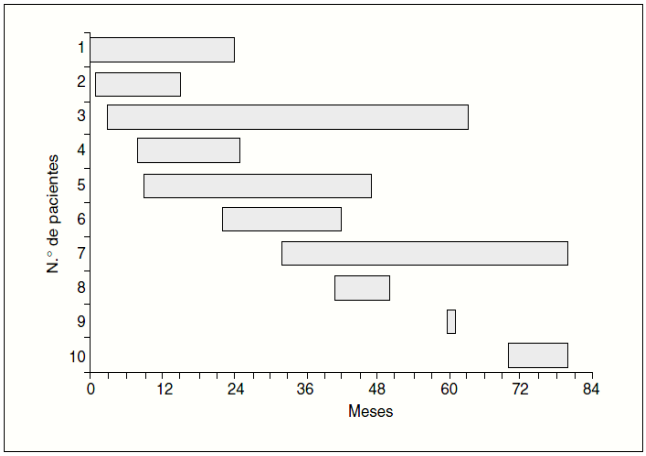


***Figura 2**. Tiempo de calendario.*

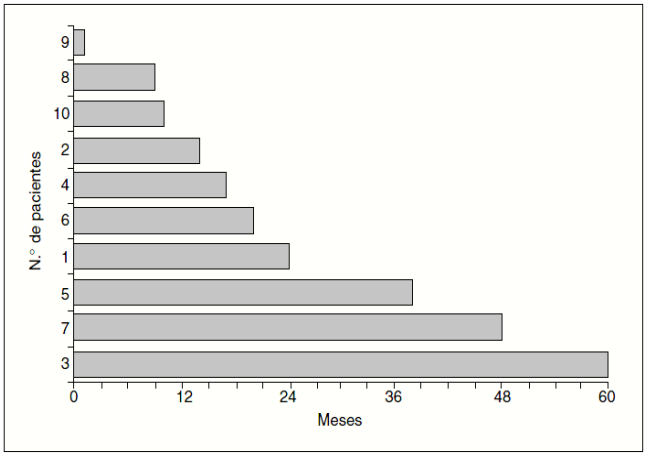


***Figura 3.** Tiempo de seguimiento.*

En la *figura 2*, se observa una muestra de individuos determinados por su tiempo de inclusión en el estudio y su fecha de observación o censura, mientras que en la *figura 3*, se han representado a tiempo inicial de estudio y se han ordenado de menor a mayor para una mejor observación gráfica de su tiempo de supervivencia.

Se denomina análisis de supervivencia al conjunto de técnicas que permiten estudiar la variable tiempo hasta que ocurre un evento y su dependencia de otras posibles variables explicativas teniendo en cuenta la información parcial contenida en las censuras (ABRAIRA, 2004); término que se va a definir a continuación.

2.1.Censuras

Las censuras se presentan en el estudio cuando no se puede determinar de una manera precisa el tiempo de ocurrencia del evento fijado, teniendo entonces, una información incompleta de ciertos individuos.

En el siguiente gráfico, se observan las censuras y eventos que se han producido en una muestra de 15 individuos tomada de la base de datos *"btrial"*, la cual, pertenece al programa estadístico R.

Los individuos representados con un círculo rojo son las censuras, mientras que los individuos a los que le ha ocurrido el evento se muestran con un triángulo.

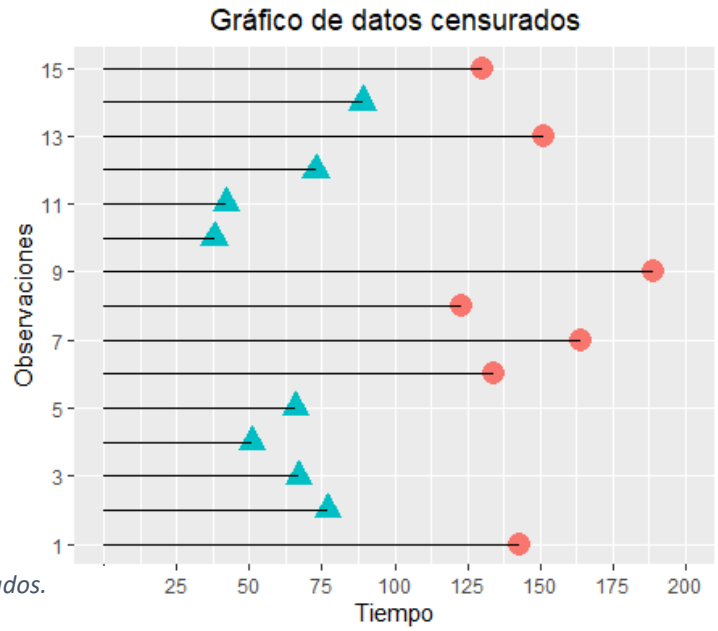


***Gráfico 1.** Gráfico de ejemplo de datos censurados.*

En análisis de supervivencia las censuras se clasifican en tres categorías relevantes: censura por la derecha, censura por la izquierda y censura por intervalos.

- Censura por la derecha

Este tipo de censura es la más común y se produce cuando el tiempo hasta el evento, T, es mayor que un tiempo de observación o censura, C, es decir, cuando el evento no se ha producido mientras el periodo de observación, ya que el suceso no se observa antes del fin de estudio.

Los datos se representan a partir de dos variables aleatorias (X, δ); siendo:

$$X = \min(T, C), \qquad \delta = \begin{cases} 1\ si\ T\ \leq C: ocurrencia\ del\ evento \\ 0\ si\ T > C: observación\ censurada \end{cases}$$

Incluidos dentro de este conjunto, se tienen tres tipos de censuras que están determinadas por la causa que ha provocado la misma:

Censura de tipo I. En este caso, se define una duración concreta de estudio y se determinará como censura cuando el tiempo de evento sea superior al tiempo de fin de estudio.

Censura de tipo II. Para este tipo de censura, se determina un número concreto de eventos que marcará el fin del estudio, es decir, si se tienen n observaciones, cuando se produzca el *r-ésimo* suceso $(r < n)$, se dará por concluido el estudio. Solamente se observan los primeros r tiempos y el resto de tiempos, $n - r$, serán observaciones censuradas por la derecha a tiempo r. A diferencia del caso anterior, el cierre del estudio está determinado por una proporción de individuos respecto del total, mientras que el *tipo I*, lo marca un tiempo. En algunas ocasiones, se fija un tiempo final cuando no se llega a la proporción buscada en un periodo largo de tiempo.

Censura de tipo III. Esta censura, también denominada no informativa o aleatoria, se presenta con un acontecimiento aleatorio durante el estudio, se puede producir en el momento en el que se pierde la información sobre el individuo, cuando le ocurre un evento diferente al prefijado o cuando aún al final del estudio, no se ha presentado el evento. Además, este tipo de censura se caracteriza por la independencia entre el tiempo hasta el evento y el tiempo de censura, por ello se denomina no informativa.

Una censura se considera informativa cuando existe dependencia entre el tiempo de suceso y el tiempo observado, es decir, cuando un individuo se ve afectado por razones relacionadas con el estudio y debe abandonarlo. Esta censura no se considerará en este trabajo.

- Censura por la izquierda

Se dice que una observación está censurada por la izquierda cuando el evento ocurre previamente a la incorporación en estudio, es decir, cuando un individuo va a ser analizado por primera vez y se tiene que el evento con el que se inicia el estudio ya se ha producido y, además se desconoce su tiempo de suceso.

Este tipo de censura es poco frecuente en análisis de supervivencia, pero en algunos estudios cabe la posibilidad de darse. Un ejemplo en el que puede aparecer este tipo de censura es en un estudio entre un infarto y otro, en el que, tras sufrir el primer infarto, el individuo entra en estudio, sabiendo que le ha ocurrido, pero desconociendo el momento exacto. Es decir, la censura por la izquierda se caracteriza por la búsqueda de un segundo instante en el que se produzca el evento, teniendo en cuenta que el caso anterior sucedió fuera de estudio.

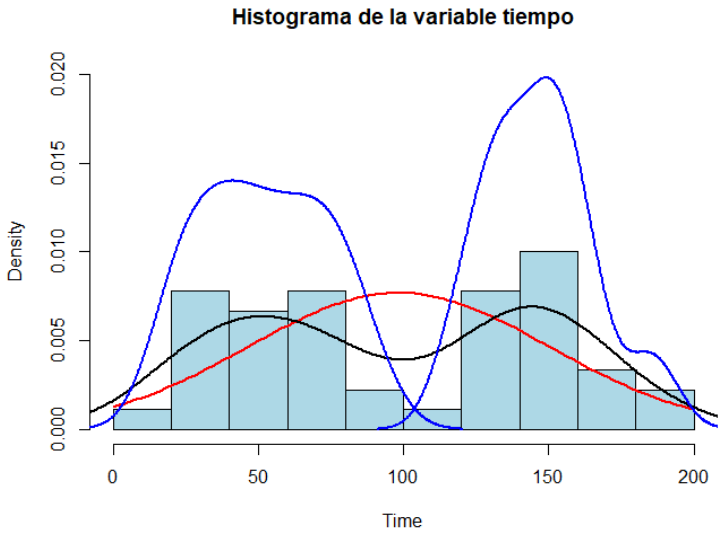
- Censura por intervalos

En este tipo de censura se desconoce el momento en el que se produce la misma, simplemente se tiene la información de que se da entre dos instantes de tiempo en un periodo de observación. Las censuras por intervalos suelen ocurrir en estudios en los que la variable dependiente está dividida en intervalos de tiempo, de manera que el evento puede darse dentro de dicho intervalo, pero se desconoce el momento exacto.

A consecuencia de la asimetría de la variable tiempo y la presencia de censuras, la distribución Normal no será la más adecuada para modelizar la variable tiempo hasta un evento; así mismo, podría no ser adecuado utilizar la media como estimador del valor de la variable. La mediana y los percentiles serán los estimadores más usados.

Además, tampoco es adecuado utilizar las técnicas de regresión habituales en este tipo de datos. En su defecto, se utilizan otros métodos, tanto paramétricos, como los que utilizan las distribuciones Exponencial o Weibull, como no paramétricos, que no precisan de distribuciones determinadas. Más adelante, se expondrán estos métodos.

En el *gráfico 2*, se observa la distribución que sigue la variable tiempo en la base de datos seleccionada, *"btrial"*. Como se puede comprobar, no siguen una distribución Normal (línea roja). Si se observa la curva de densidad de la variable, se pueden dividir los datos en dos grupos, formando uno con los valores que toma la variable inferiores a 100 y, el otro, con valores superiores o iguales a 100; con ello, se observa que cada uno de estos conjuntos sigue una distribución Normal. Esto se comprueba con el test de normalidad de Shapiro-Wilk, donde para la variable tiempo se obtiene un p-valor inferior a 0.05 y, por lo tanto, se concluye el rechazo de la hipótesis nula de normalidad; mientras que para las otras dos divisiones de la variable principal, se obtienen p-valores muy por encima de 0.05, por lo que, no se puede rechazar el supuesto de normalidad.




```
        Shapiro-Wilk normality test

data:  Time
W = 0.91949, p-value = 0.004063
```

```
        Shapiro-Wilk normality test

data:  Time1
W = 0.9552, p-value = 0.3735
```

```
        Shapiro-Wilk normality test

data:  Time2
W = 0.97256, p-value = 0.7695
```

Gráfico 2. *Gráfico de densidad de la variable tiempo ("btrial").*

2.2. Funciones relevantes en análisis de supervivencia

Las funciones específicas del análisis de supervivencia son la función de densidad, la función de distribución, la función de supervivencia y la función de riesgo, junto con la de riesgo acumulado.

Sea T una variable aleatoria no negativa que denota el tiempo de supervivencia o tiempo hasta que ocurre el evento. Se presentan aquí las diferentes funciones asociadas a esta variable dependiendo de la distribución que siga.

Cuando la variable T es discreta, en la mayoría de las ocasiones el tiempo está estructurado en intervalos o toma valores enteros positivos, t_j con $j = 1, 2, \dots$.

2.2.1. Función de densidad de probabilidad y función de distribución

La función de densidad de probabilidad de la variable T y sus propiedades, para el caso discreto de la variable es:

$$\circ\ f(t_j) = P(T = t_j)\ con\ j = 1, 2 \dots y\ \ t_1 < t_2 < t_3 < \ \dots \quad \circ \sum_{t_j \in T} f(t_j) = 1\ y\ f(t_j) \geq 0$$

Mientras que para el caso continuo se tendría:

$$\circ\ f(x)\colon \mathbb{R} \to \mathbb{R}\ tal\ que\ P(X \in (a, b)) = \int_a^b f(x)dx \quad \circ \int_{\infty}^{\infty} f(x)dx = 1\ con\ f(x) \geq 0$$

Esta función de densidad determina la probabilidad instantánea de que el evento de interés se produzca en un tiempo t.

Por otro lado, la función de distribución para ambos casos queda definida por las siguientes expresiones, respectivamente:

$$\circ\ F(t) = P(T \leq t) = \sum_{t_j \leq t} f(t_j) \qquad \circ\ F(t) = P(T \leq t) = \int_0^t f(x)dx$$

2.2.2. Función de supervivencia

La probabilidad de supervivencia se define como la probabilidad de permanecer en el estudio sin que se produzca el evento sobre los individuos. Por lo tanto, se tiene que la función de supervivencia, o tasa acumulada, es la probabilidad de que un individuo en el estudio no presente el evento antes de un tiempo determinado; es decir, que sobreviva hasta un tiempo concreto t.

Sea T una variable aleatoria y positiva con una función de distribución $F(t)$ y función de densidad de probabilidad $f(t)$; se tiene que la función de supervivencia, $S(t)$, está definida por

$$S(t) \ = \ 1 - F(t) \ = P(T > t)$$

Esta función, asocia a cada tiempo t la probabilidad de que un sujeto sobreviva a dicho instante de tiempo. Además, se puede destacar que la función en $S(0) = 1$, es decir, que un individuo siempre está vivo al comienzo del estudio, mientras que para un tiempo infinito, dicha probabilidad es nula; es decir, $S(t) = 0$ cuando $t \to \infty$. Por lo tanto, como la función de distribución es creciente hacia 1, se observa que la función de supervivencia es una función decreciente.

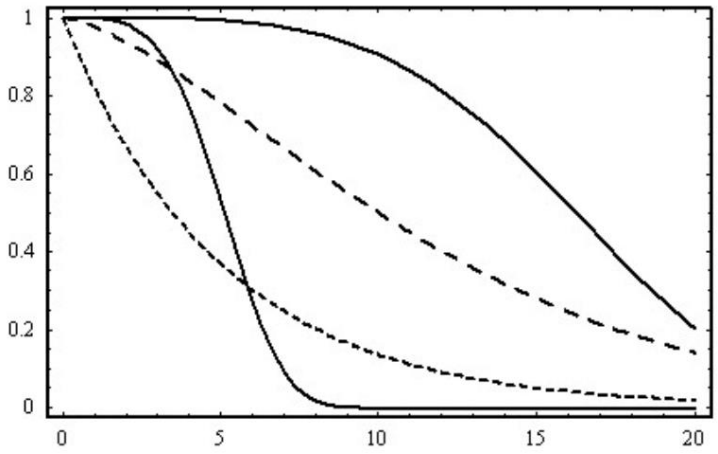


***Figura 4**. Distintas funciones de supervivencia.*

Si T es una variable aleatoria discreta que toma valores t_j con $j = 1, 2, \ldots$ y función de probabilidad $f(t_j) = P(T = t_j)$ donde $t_1 < t_2 < \ldots$; la función de supervivencia será:

$$S(t) \ = \ P(T > t) \ = \sum_{t_j > t} f(t_j)$$

De la misma manera, cuando T es una variable aleatoria continua, su función de supervivencia, es la siguiente integral de la función de densidad.

$$S(t) \;=\; P(T > t) \;=\; \int_t^{\infty} f(x)\,dx$$

2.2.3. Función de riesgo y función de riesgo acumulado

La función de riesgo, $h(t)$, representa la evolución de la probabilidad de evento en relación con el tiempo de supervivencia de los individuos.

Esta función, también denominada tasa condicionada de mortalidad (*Hazard rate*), representa el número de casos que presentan el evento en un momento determinado entre el número de casos que llegan a dicho momento sin haberlo experimentado.

Cuando la variable T es discreta, la función de riesgo determina la probabilidad condicional de que el evento ocurra a tiempo $T = t_j$, condicionada a que el sujeto permanece vivo antes de dicho tiempo, luego se define como

$$h(t_j) = P(T = t_j \mid T \geq\ t_j) = \frac{P(T = t_j)}{P(T \geq t_j)} = \frac{f(t_j)}{S(t_j)}$$

Si T es una variable continua, la función de riesgo se define como la probabilidad de que el tiempo de evento se produzca en un intervalo de tiempo pequeño, suponiendo la supervivencia del individuo al inicio de dicho intervalo; es decir, la probabilidad condicional de que le ocurra el evento a un individuo en un tiempo reducido, de t a $t + \Delta t$, sabiendo que ha sobrevivido hasta el tiempo t. Su fórmula sería:

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

En líneas generales, es la probabilidad de que un individuo que está siendo observado a tiempo t le suceda el evento en ese momento, por ello, esta función es la más adecuada para la descripción del estudio, porque los valores constituyen las tasas de incidencia del evento analizado.

Esta función es importante en análisis de supervivencia, ya que facilita la elección de un modelo paramétrico, como el Exponencial o Weibull, según la forma que tome en cada situación.

La función de riesgo puede ser creciente, decreciente, constante o tener forma de bañera. Cada una de estas formas viene representada por distribuciones diferentes. Para un riesgo creciente se habla de poblaciones que envejecen, mientras que si se tiene una función de riesgo decreciente hablamos de individuos que mejoran con el paso del tiempo. Una función de riesgo que permanece constante proporciona una tasa de riesgo que no varía durante todo el tiempo de observación.

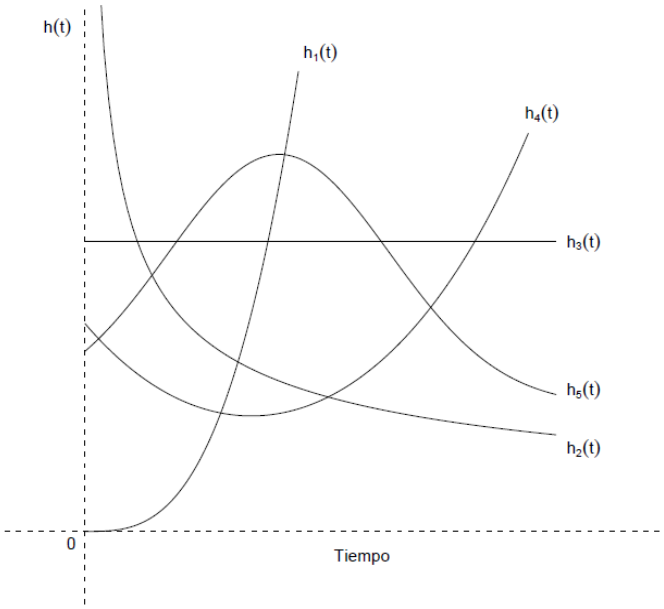


Figura 5. *Distintas funciones de riesgo.*

Por ejemplo, si se dispone de una función de riesgo constante, puede proceder de un conjunto de individuos jóvenes sanos según el efecto de los accidentes laborales; si se observa una curva creciente, tal vez pueda pertenecer a pacientes que se encuentran en tratamiento, el cual, se comprueba con el tiempo, que no funciona; para una situación de riesgo decreciente se puede observar una población joven que padece cierta enfermedad pero tiene cura mediante una operación, el comienzo tras la operación tendrá ciertos riesgos, pero a medida que avanza el tiempo éste disminuirá; y, por último, un ejemplo para un riesgo de bañera puede ser, la población nacida en la década de los setenta, donde muchos de los nacimientos eran en muertes, en la etapa joven disminuía el riesgo y en una edad avanzada predominaban los fallecimientos.

Pese a existir el riesgo de que se produzca el evento en cualquier momento del estudio, éste será mayor al comienzo, ya que el número de individuos expuestos al riesgo es superior.

En la función de riesgo, a medida que avanza el tiempo, el intervalo de confianza aumenta su tamaño, esto se debe a la evidencia del menor número de sujetos que perduran en el estudio a su término.

A partir de esta función, se puede calcular la función de riesgo acumulado, $H(t)$, una función no decreciente que acumula el riesgo a lo largo del tiempo. Para el caso discreto se define como

$$H(t) = \sum_{t_j \leq t} h(t_j)$$

Por otro lado, para el caso continuo se tiene la expresión

$$H(t) = \int_0^t h(x)\, dx$$

Dadas las expresiones anteriores, tanto para variables discretas como para variables continuas, se pueden obtener las siguientes relaciones matemáticas entre las mismas:

- Caso discreto:

$$f(t_j) = S(t_j) - S(t_{j+1}) = h(t_j) \cdot S(t_j)$$

$$h(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)} \rightarrow 1 - h(t_j) = \frac{S(t_{j+1})}{S(t_j)}$$

$$S(t) = \prod_{t_j<t} (1 - h(t_j))$$

- Caso continuo:

$$f(t) = -\frac{d}{dt} S(t) = h(t) \cdot S(t)$$

$$h(t) = -\frac{d}{dt} \log S(t)$$

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(x)\,dx\right] \rightarrow H(t) = -\ln S(t)$$

Estas igualdades entre las funciones facilitan la obtención de cualquiera de ellas a partir de una única expresión.

2.3. Distribuciones útiles en análisis de supervivencia

Existen varios modelos paramétricos que se adecúan al análisis de supervivencia, como por ejemplo, Exponencial, Weibull, Gamma, Log-normal o Log-logística. En este caso, el trabajo se centra en los modelos mencionados inicialmente, es decir, el modelo Exponencial y el modelo Weibull. Esto no quiere decir que éstas sean siempre las mejores distribuciones para un estudio de análisis de supervivencia, ya que habría que hacer un estudio previo para obtener el modelo que mejor se ajuste a los datos dados.

La preferencia por optar por un modelo u otro viene determinado por la forma de su función de riesgo, ya que, a través de ella, se observa una aproximación de la información que determina la causa del evento, estableciendo así, una curva que describe la tasa de riesgo que evoluciona en el tiempo. Por ejemplo, si se somete a un paciente a una operación, el riesgo aumentará en los instantes posteriores, pero si sobrevive, disminuirá hasta estabilizarse. Por lo tanto, se tendría una función de riesgo a modelar creciente en los primeros tiempos, que toma un punto máximo y después, va disminuyendo a lo largo del tiempo hasta llegar a mantenerse constante.

2.3.1. Modelo Exponencial

Se puede suponer una distribución exponencial, $\mathcal{E}(\lambda)$, de parámetro positivo λ, cuando la función de riesgo se comporta de manera constante, es decir, el riesgo de que suceda el evento no se modifica con el paso del tiempo.

Por lo tanto, se define como

$$h(t) = \lambda$$

Esta propiedad, denominada como *pérdida o ausencia de memoria*, hace que la probabilidad de que suceda el evento en un tiempo t, no dependa de ningún tiempo anterior, es decir, la probabilidad de evento a tiempo t es independiente de la probabilidad de evento a tiempos $t-1, t-2, \ldots$.

A partir de dicha función, se obtienen el resto de funciones características del análisis de supervivencia:

$$H(t) = \int_0^t h(t)dt = \lambda t$$

$$S(t) = e^{-H} = e^{-\lambda t}$$

$$F(t) = 1 - S(t) = 1 - e^{-\lambda t}$$

$$f(t) = F'(t) = -e^{-\lambda t} \cdot -\lambda = \lambda e^{-\lambda t}$$

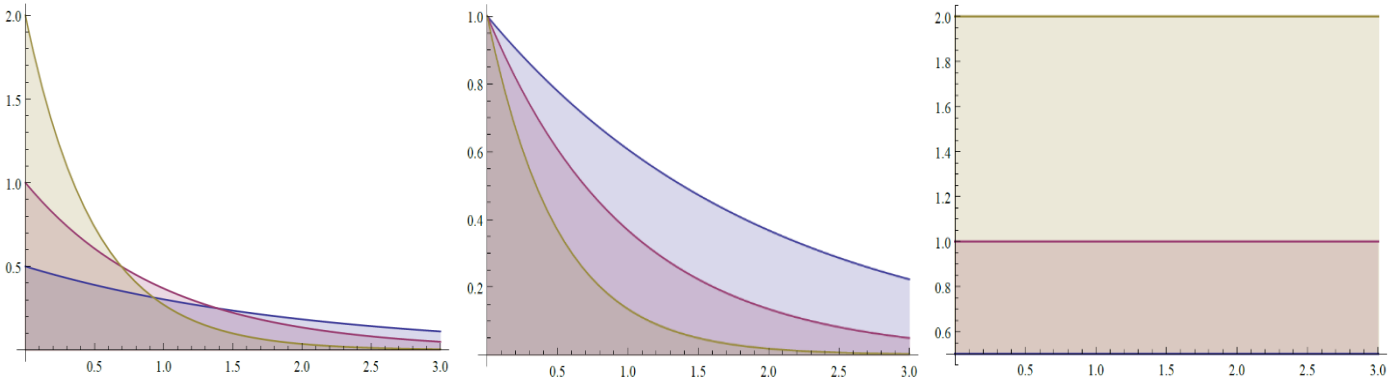


Figura 6. *Funciones características de la distribución Exponencial (densidad, supervivencia y riesgo).*

La media de la distribución exponencial es $1/\lambda$ y su varianza $1/\lambda^2$. Como esta distribución está sesgada a la derecha, una mejor estimación se tomaría a partir de la mediana, determinada por $\ln 2\ /\lambda$.

2.3.2. Modelo Weibull

La distribución de Weibull está determinada por dos parámetros positivos, γ y λ, denominados parámetros de forma y escala, respectivamente.

A consecuencia de la cantidad de valores que puede tomar el parámetro de forma, obteniendo así gran variedad de curvas, esta distribución es una de las más utilizadas en análisis de supervivencia.

Como se ha mencionado anteriormente, la función de riesgo varía con el tiempo y puede tener una forma aleatoria. Se expresa de la siguiente forma:

$$h(t) = \gamma\lambda t^{\gamma-1}$$

En el caso de que el parámetro $\gamma > 1$ la función será creciente y decreciente en el caso contrario, $\gamma < 1$. Cuando $\gamma = 1$, coincide con el modelo $\mathcal{E}(\lambda)$ y la función de riesgo sería constante. Estas distintas formas que puede tomar la función de riesgo favorecen la posible modelación del tiempo.

El resto de funciones de esta distribución vienen determinadas por las siguientes expresiones:

$$H(t) = \int_0^t h(t)dt = \lambda t^{\gamma}$$

$$S(t) = e^{-H} = e^{-\lambda t^{\gamma}}$$

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

$$f(t) = F'(t) = -e^{-\lambda t^{\gamma}} \cdot (-\lambda\gamma t^{\gamma-1}) = (\lambda\gamma t^{\gamma-1}) \cdot e^{-\lambda t^{\gamma}}$$

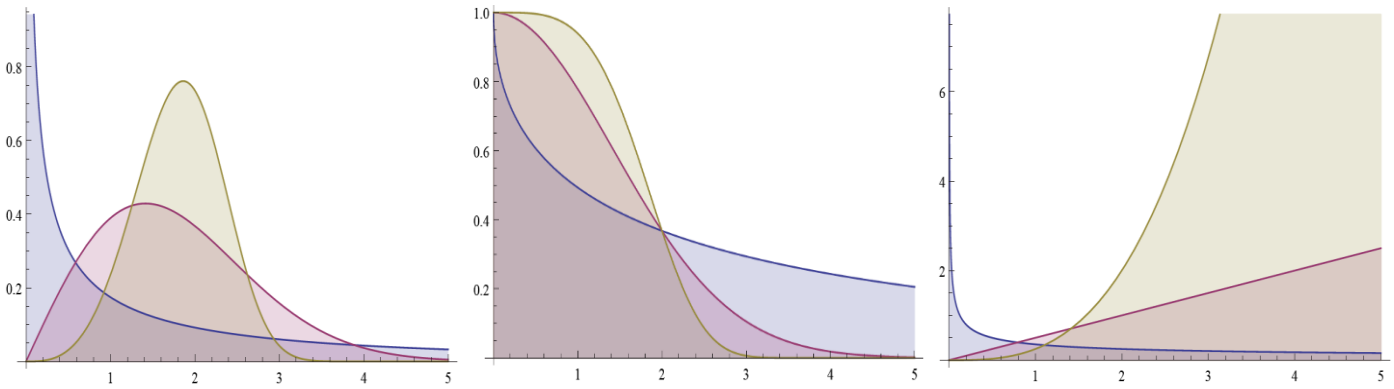


Figura 7. *Funciones características de la distribución Weibull (densidad, supervivencia y riesgo).*

Su media y varianza están definidas por las siguientes fórmulas

$$E(T) = \frac{\Gamma\left(1+\frac{1}{\gamma}\right)}{\lambda^{\gamma}} \qquad V(T) = \frac{\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)}{\lambda^{2/\gamma}}$$

[14]

Donde $\Gamma(t)$ es la función gamma de Euler, que se define como la integral para $t > 0$,

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy$$

Normalmente, es más común tomar la mediana como estimador, ya que es una medida de tendencia central y la distribución Weibull se caracteriza por ser sesgada a la derecha. Su fórmula es

$$Me(T) = \left(\frac{\log 2}{\lambda}\right)^{1/\gamma}$$

3. Análisis no paramétrico de la supervivencia

En análisis de supervivencia, debido al problema de la presencia de censuras, unos estadísticos serán más utilizados que otros, ya que, por ejemplo la media, provocaría una estimación sesgada de la muestra. Los estimadores que cumplan las propiedades estadísticas necesarias serán aquellos más utilizados, como la mediana o la varianza.

A consecuencia de esta dificultad, se precisan otros métodos para la estimación de la supervivencia. La supervivencia o el riesgo determinan la distribución que siguen los datos observados, por ello, uno de nuestros objetivos será la estimación de los parámetros involucrados en dicha distribución. Generalmente, las estimaciones se realizan a partir de la función de supervivencia y la función de riesgo.

El estimador más utilizado en análisis de supervivencia es el de Kaplan-Meier, ya que permite trabajar fácilmente con datos censurados. Es un estimador de la función de supervivencia y para poder realizar su cálculo se deben tener los tiempos de supervivencia de cada sujeto, que, por lo general, se suelen tener. Como se tienen en cuenta los tiempos de supervivencia individuales, se dice que es un estimador bastante preciso.

Además, el análisis no paramétrico permite realizar comparaciones de las curvas de supervivencia de dos o más grupos.

3.1. Método producto límite de Kaplan-Meier

El método de Kaplan-Meier o método del producto límite, calcula la probabilidad de supervivencia para cada tiempo t_j en el que se produce un evento, con $j = 1, 2, \dots, r$ tiempos de evento. Por lo que con él, se obtiene una estimación de la función de supervivencia.

El tiempo transcurrido entre cada evento está marcado por un intervalo I_j, que transcurre desde un tiempo t_{j-1} hasta el tiempo de evento t_j, $I_j = (t_{j-1}, t_j]$. Para cada tiempo de evento t_j, se definen, por un lado, como r_j, el número de individuos, tales que, cumplen la condición $T > t_{j-1}$, es decir, aquellos que llegan vivos al intervalo I_j; y, por otro lado, d_j como el número de individuos que experimenta el evento en dicho intervalo.

Por lo tanto, se tiene que la probabilidad de supervivencia a tiempo t_j es la probabilidad de que el evento se produzca después de dicho tiempo t_j. Utilizando el teorema de la probabilidad compuesta (Regla del producto) se tiene que:

$$P(T > t_j) = P(T > t_{j-1}) \cdot P(T > t_j | T > t_{j-1}) \rightarrow S(t_j) = S(t_{j-1}) \cdot \frac{r_j - d_j}{r_j}$$

Esta probabilidad relaciona la supervivencia a tiempo t_j con la supervivencia en un instante previo, t_{j-1}.

La probabilidad condicionada de sobrevivir a un tiempo t_j habiendo sobrevivido a un tiempo anterior, se denomina probabilidad de supervivencia y está determinada por r_j y d_j.

Esta probabilidad solamente se calculará en instantes de tiempo en que se produzca el evento y, en otros casos, se mantendrá la del tiempo anterior. Esta tasa se define con la siguiente expresión:

$$Probabilidad\ de\ supervivencia = \frac{r_j - d_j}{r_j} = 1 - \frac{d_j}{r_j}$$

Esta probabilidad, aplicada de manera sucesiva para cada intervalo de tiempo, será el producto de la sucesión de la probabilidad de supervivencia para todos los tiempos t_j y proporcionará una estimación de la curva de supervivencia:

$$\hat{S}(t) = \prod_{t_j \leq t} \frac{r_j - d_j}{r_j}$$

Además, cabe destacar que el estimador de Kaplan-Meier es un estimador máximo verosímil de la función de supervivencia y cumple con las propiedades de insesgadez, consistencia y eficiencia de un estimador.

Para proceder al cálculo del estimador, se deben ordenar los tiempos de supervivencia de los individuos de menor a mayor, teniendo en cuenta los empates de tiempos de supervivencia entre individuos. Una vez ordenados, se calcula para cada tiempo la probabilidad de supervivencia, expresión mencionada anteriormente, teniendo en cuenta el número de individuos en riesgo y el número de eventos ocurridos para cada intervalo I_j. Se aplica sucesivamente dicha fórmula para todos los tiempos y finalmente, se calcula el estimador de K-M, para obtener la curva de supervivencia de la población en estudio.

Esta curva de supervivencia es una estimación de la función de supervivencia y representa una función decreciente, constante entre dos tiempos de evento consecutivos y decreciente en los tiempos en que se produce el suceso. Se caracteriza por valer 1 antes del primer tiempo de evento y tender a cero a medida que el tiempo de estudio avanza, aunque no siempre llega a dicho valor, por ejemplo cuando se encuentran censuras por fin de estudio.

La forma de la curva de supervivencia revela la velocidad con que se va dando el evento según evoluciona el tiempo.

A partir de este estimador, también se puede obtener una estimación de la función de riesgo acumulado, que se relaciona con la función de supervivencia mediante la expresión $H(t) = -\ln S(t)$. Por lo tanto, mediante el estimador de Kaplan-Meier, se tiene la estimación:

$$\widehat{H}(t) = -\ln \hat{S}(t)$$

A continuación, se va a realizar un ejemplo para el cálculo del estimador con una muestra aleatoria, tomada de la base de datos anteriormente utilizada (“*btrial*”). Se han tomado 15 individuos de dicha muestra para mostrar el procedimiento para una muestra pequeña de datos.

La muestra obtenida de los tiempos de evento es la siguiente: 189+, 38, 89, 66, 73, 51, 42, 151+, 143+, 164+, 134+, 123+, 77, 67 y 130+, marcados con un símbolo + aquellos individuos que han sido censurados. Con la tabla siguiente se resume el cálculo del estimador descrito anteriormente:

Tabla 1. *Tabla del estimador de Kaplan-Meier para una muestra de datos.*

Tiempo de evento	Individuos en riesgo	Número de eventos	Tasa de supervivencia	Estimador K-M
38	15	1	$(15-1)/15 = 0.933$	$14/15 = 0.933$
42	14	1	$(14-1)/14 = 0.929$	$14/15 \cdot 13/14 = 0.867$
51	13	1	$(13-1)/13 = 0.923$	$14/15 \cdot 13/14 \cdot 12/13 = 0.800$
66	12	1	$(12-1)/12 = 0.917$	$14/15 \cdot 13/14 \cdot 12/13 \cdot 11/12 = 0.733$
67	11	1	$(11-1)/11 = 0.909$	$14/15 \cdot 13/14 \cdot 12/13 \cdot 11/12 \cdot 10/11 = 0.667$
73	10	1	$(10-1)/10 = 0.900$	$14/15 \cdot 13/14 \cdot 12/13 \cdot 11/12 \cdot 10/11 \cdot 9/10 = 0.600$
77	9	1	$(9-1)/9 = 0.889$	$14/15 \cdot 13/14 \cdot 12/13 \cdot 11/12 \cdot 10/11 \cdot 9/10 \cdot 8/9 = 0.533$
89	8	1	$(8-1)/8 = 0.875$	$14/15 \cdot 13/14 \cdot 12/13 \cdot 11/12 \cdot 10/11 \cdot 9/10 \cdot 8/9 \cdot 7/8 = 0.467$

Se tiene que, a medida que pasa el tiempo, la estimación de la supervivencia va disminuyendo en los tiempos en los que se producen eventos. Los tiempos de evento superiores a los 89 días no se reflejan en la tabla, esto se debe a que los individuos que sobreviven a dicho tiempo son censuras y, por lo tanto, su supervivencia se marcaría como la del último evento.

Con estos resultados, se puede obtener la curva de supervivencia. La gráfica que la representa es la siguiente:

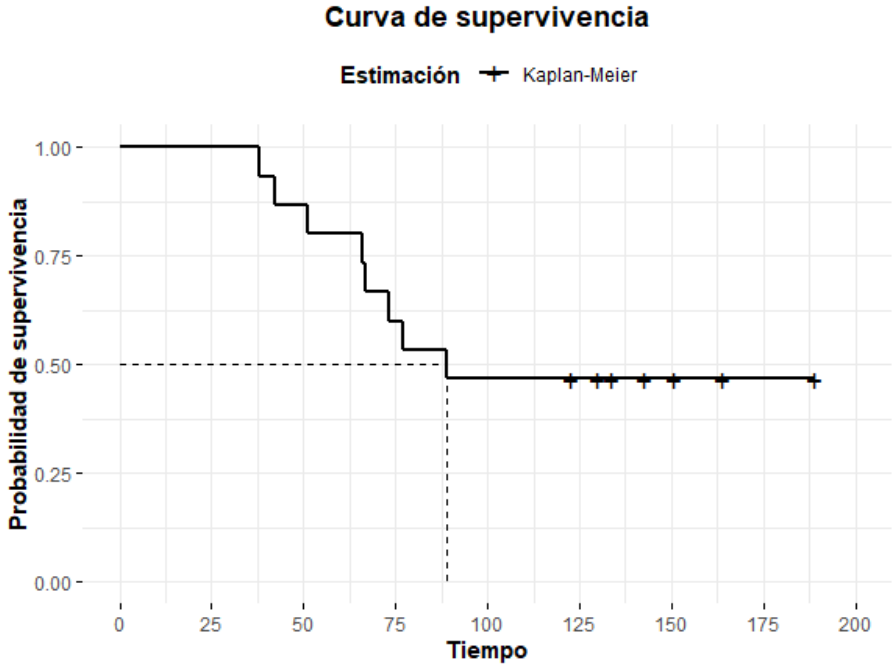


Gráfico 3. *Curva de supervivencia para la base de datos "btrial".*

Mediante la curva de supervivencia, se puede determinar la mediana del tiempo de supervivencia, y además, la supervivencia de los individuos para cualquier tiempo t_j, aunque una interpretación más adecuada sería observarla para toda la duración del estudio.

Se observa que la supervivencia en los tiempos iniciales del estudio comienza a disminuir a partir de los 38 meses hasta los 89, tiempo que marca la mediana del tiempo de estudio. Después de dicho tiempo, no se producen más eventos y simplemente se observan censuras, tiempos marcados con una cruz sobre la curva, lo que provoca que la estimación de la función de supervivencia esté indefinida. Por lo tanto, se podría concluir, que una vez superados los 89 meses, los individuos tendrán una supervivencia del 46.7%.

A partir de este estimador, también se puede realizar una comparación entre dos o más grupos. De la misma manera que en el caso anterior, se realizan los cálculos necesarios, en este caso para dos grupos (Respuesta inmunohistoquímica negativa o positiva), por lo tanto, para la comparación de dos curvas de supervivencia. Los resultados son los siguientes:

Tabla 2. *Tabla del estimador Kaplan-Meier para grupo de respuesta inmunohistoquímica negativa.*

Tiempo de evento	Individuos en riesgo	Número de eventos	Tasa de supervivencia	Estimador K-M
51	10	1	$9/10 = 0.90$	$9/10 = 0.90$
66	9	1	$8/9 = 0.889$	$9/10 \cdot 8/9 = 0.80$
67	8	1	$7/8 = 0.875$	$9/10 \cdot 8/9 \cdot 7/8 = 0.70$

Tabla 3. *Tabla del estimador Kaplan-Meier para grupo de respuesta inmunohistoquímica positiva.*

Tiempo de evento	Individuos en riesgo	Número de eventos	Tasa de supervivencia	Estimador K-M
38	5	1	$4/5 = 0.800$	$4/5 = 0.800$
42	4	1	$3/4 = 0.750$	$4/5 \cdot 3/4 = 0.600$
73	3	1	$2/3 = 0.667$	$4/5 \cdot 3/4 \cdot 2/3 = 0.400$
77	2	1	$1/2 = 0.500$	$4/5 \cdot 3/4 \cdot 2/3 \cdot 1/2 = 0.200$
89	1	1	$0/1 = 0.00$	$4/5 \cdot 3/4 \cdot 2/3 \cdot 1/2 \cdot 0/1 = 0.00$

Al dividir a los individuos en grupos según su respuesta inmunohistoquímica, la supervivencia difiere entre sus categorías, negativa y positiva. Si se observan ambas tablas, el grupo negativo obtiene una supervivencia del 70%, mientras que si la respuesta es positiva, a final del estudio, ningún individuo sobrevivió. Los gráficos correspondientes a las curvas de supervivencia de cada grupo se encuentran en el ANEXO Esquema de un análisis de supervivencia.(GRÁFICO 14. CURVA DE SUPERVIVENCIA PARA UNA RESPUESTA INMUNOHISTOQUÍMICA NEGATIVA. y GRÁFICO 15. CURVA DE SUPERVIVENCIA PARA UNA RESPUESTA INMUNOHISTOQUÍMICA POSITIVA.**Gráfico 15**).

Con estas diferencias observables a simple vista, se puede concluir que hay disparidad entre los grupos, pero existen test estadísticos que, mediante pruebas no paramétricas, demuestran dicha disimilitud.

3.2.Comparación de curvas de supervivencia

Los gráficos de las curvas de supervivencia de Kaplan-Meier muestran una idea de la diferencia entre las curvas de supervivencia de dos o más grupos, pero se necesita una prueba estadística con la que afirmar una diferencia significativa cierta.

Existen varios métodos para contrastar la igualdad de las funciones de supervivencia entre diferentes grupos. Si se diera el caso de no tener observaciones censuradas en el conjunto de datos, el test estadístico más utilizado es el de la suma de rangos de Wilcoxon, pero en la mayoría de las ocasiones se trabajará con censuras.

La prueba no paramétrica más importante en el caso de trabajo con censuras es el test *Log-rank*, que se describirá a continuación, junto con otras pruebas no paramétricas también utilizadas pero menos destacadas. Para todos los test, se tiene la hipótesis nula de igualdad de supervivencia entre los grupos.

3.2.1. Test Log-rank o prueba de Mantel-Haenszel

El test log-rank se basa en el cálculo del número de sucesos que se esperan para cada tiempo, asumiendo la igualdad de las funciones de supervivencia entre los grupos. El test compara el número de eventos observados en cada grupo con el número de eventos esperados y considera la evolución total de la curva. Con cada evento se calcula el número observado y esperado de eventos en cada grupo, asumiendo que no hay diferencias entre ellos, es decir, considerando H_0 cierta.

La premisa de la que parte este test es de la similitud entre el número de eventos observados en los distintos intervalos de tiempo para la comparación de los grupos. Aunque en la mayoría de las ocasiones no se de este suceso, lo que la hipótesis nula afirma es que, globalmente, la supervivencia en ambos grupos sea semejante.

Para llevar a cabo este test, se ordenan cronológicamente los individuos según el tiempo de evento t_j de ambos grupos, considerándolos un único conjunto y formando intervalos de la forma $I_j = (t_{j-1}, t_j]$. A continuación, se determinan el número de sujetos en riesgo r_j para cada intervalo, el número de eventos observados y el número de eventos esperados para cada grupo y en total. Posteriormente, se obtiene el total de eventos tanto observados como esperados para cada grupo y se calcula el estadístico. El número total de pérdidas observadas para cada grupo i se denotará como O_i y el número total de pérdidas esperadas para cada grupo i como E_i. El método es válido para comparar dos o más grupos de observaciones.

El estadístico obtenido, denominado en este trabajo LR, sigue un distribución chi-cuadrado con $k-1$ grados de libertad, siendo k el número de grupos o curvas de supervivencia a comparar. Éste se define como

$$LR = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

Este estadístico resultante se compara con la distribución chi-cuadrado y con el p-valor obtenido se puede decidir la existencia o no de diferencias estadísticas entre las curvas de supervivencia. En el caso de trabajar con un tamaño de muestra grande, el estadístico puede aproximarse a una normal tipificada.

Si continuamos analizando la muestra de la base de datos "*btrial*" realizando el test *log-rank* para los dos grupos formados por la variable respuesta inmunohistoquímica en el que el valor 1 que toma la variable será la respuesta negativa y, en el otro caso, el valor 2, equivale a una respuesta positiva. Se obtienen los siguientes resultados:

Tabla 4. *Tabla del test log-rank para la muestra de la base de datos "btrial".*

Tiempo de evento	Pacientes en riesgo			Pérdidas observadas			Pérdidas esperadas		
	IM = 1	IM = 2	Total	IM = 1	IM = 2	Total	IM = 1	IM = 2	Total
38	10	5	15	0	1	1	0.67	0.33	1.00
42	10	4	14	0	1	1	0.71	0.29	1.00
51	10	3	13	1	0	1	0.77	0.23	1.00
66	9	3	12	1	0	1	0.75	0.25	1.00
67	8	3	11	1	0	1	0.73	0.27	1.00
73	7	3	10	0	1	1	0.70	0.30	1.00
77	7	2	9	0	1	1	0.78	0.22	1.00
89	7	1	8	0	1	1	0.88	0.13	1.00
			Total	3	5	8	5.98	2.02	8

Aplicando la fórmula LR anteriormente mencionada, se obtiene un valor del estadístico de 5.8826, que, comparado con la distribución chi-cuadrado bilateral con 1 grado de libertad y un p-valor de 0.05, con un valor crítico inferior de $\chi^2_{1-\alpha/2} = 0.0010$ y un valor crítico superior de $\chi^2_{\alpha/2} = 5.0239$, se tiene que el estadístico no se incluye en dicha región de aceptación, por lo tanto, se concluye el rechazo de H_0, y con ello, la igualdad de las curvas de supervivencia.

Cabe destacar que el grupo de respuesta inmunohistoquímica positiva tendrá una peor supervivencia, ya que el número observado de pérdidas es mayor que el número esperado de las mismas. Esta medida se refleja en el cálculo del riesgo relativo para cada grupo y del odds ratio para la comparación entre ellos.

El cociente de los riesgos (Hazard ratio) entre ambos grupos, pérdidas observadas entre pérdidas esperadas para cada grupo, cuantifica la diferencia entre la supervivencia de un grupo respecto del otro.

$$OR = \frac{O_2/E_2}{O_1/E_1} = 4.934$$

El riesgo del grupo de respuesta positiva frente al grupo de respuesta negativa tiene como resultado 4.9, es decir, los individuos con respuestas inmunohistoquímicas negativas sobreviven 4.9 veces más que en los pacientes con respuestas positivas.

Si se realiza este test mediante el software *RStudio* los resultados que se obtienen por pantalla son los siguientes:

```
survdiff(formula = Surv(time, death) ~ im, data = btrial)

      N Observed Expected (O-E)^2/E (O-E)^2/V
im=1 10        3     5.98      1.49      5.99
im=2  5        5     2.02      4.40      5.99

 Chisq= 6  on 1 degrees of freedom, p= 0.01
```

El resultado del estadístico obtenido es de $LR = 6$ con 1 grado de libertad y un p-valor igual a 0.01, por lo que, de la misma manera que antes, como es menor que 0.05, se rechaza la hipótesis nula y se concluye que las curvas de supervivencia para los grupos de respuesta inmunohistoquímica son distintas.

Gráficamente, también se observa una diferencia clara entre las curvas. Para una respuesta negativa se tiene que la curva no llega a disminuir más allá de un 70% de supervivencia, mientras que el grupo positivo decrece muy rápidamente, llegando antes de los 100 meses de estudio a valer cero. Como se refleja en el gráfico, se muestra el p-valor de 0.014 que determina la diferencia entre las curvas.

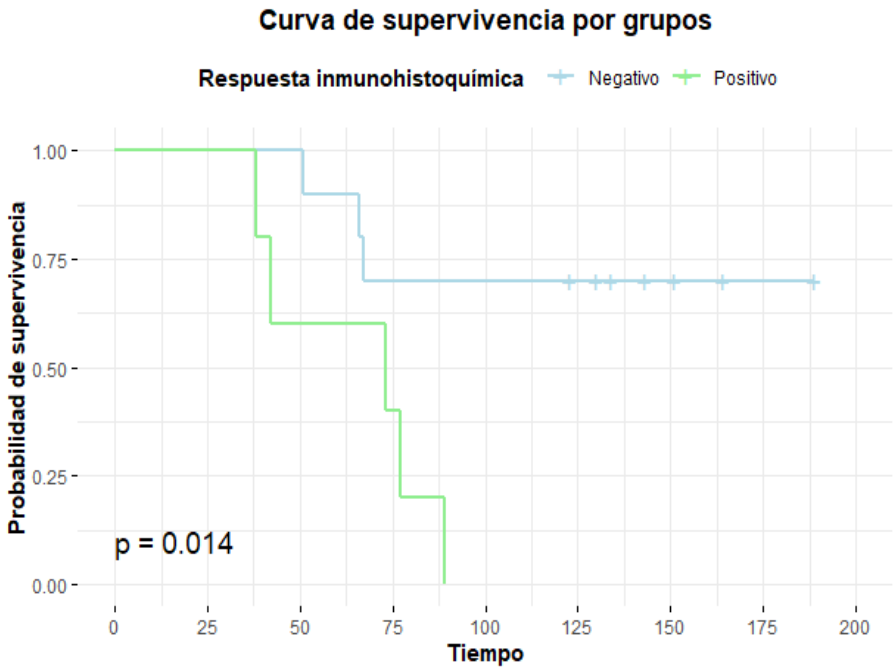


Gráfico 4. *Curva de supervivencia por grupos de respuesta inmunohistoquímica (test log-rank).*

3.2.2. Test de Wilcoxon generalizado o prueba de Breslow

Una alternativa al test Log-rank puede ser el test de Breslow o de Wilcoxon generalizado, que se basa en la suma ponderada de las diferencias observadas y esperadas entre el número de eventos. Los pesos utilizados equivalen al número de individuos en riesgo para cada tiempo t_j, es decir, la ponderación sería r_j.

Realizando el desarrollo para dos grupos $i = 1, 2$ y teniendo en cuenta la igualdad de supervivencia en ellos, se tiene que, el número esperado de eventos para el primer grupo a tiempo t_j es $e_{1j} = r_{1j}d_j/r_j$. Se define el estadístico de prueba como

$$w = \sum_{j=1}^{r} r_j(d_{1j} - e_{1j}) = \sum_{j=1}^{r} r_j\left(d_{1j} - r_{1j}\frac{d_j}{r_j}\right)$$

Con varianza igual a

$$\widehat{Var}(w) = \sum_{j=1}^{r} \frac{r_{1j}r_{2j}d_j(r_j - d_j)}{r_j^2(r_j - 1)}$$

Por lo tanto el estadístico para la prueba de Wilcoxon generalizado o de Breslow viene dado por

$$W = \frac{w^2}{\widehat{Var}(w)} \sim \chi_1^2$$

El uso del número de individuos en riesgo como ponderación hace que le de menor peso a los tiempos de supervivencia mayores, es decir, que los eventos que se produzcan al comienzo del estudio reciban mayor ponderación que los que se produzcan más adelante, ya que hay más individuos en riesgo.

Con este test, se pretende valorar la importancia de los individuos al inicio del estudio, ya que hay un mayor número de ellos en riesgo, frente a su finalización. Por ello, se dice que es un estadístico menos sensible que el anterior en la cola de la distribución de los tiempos de supervivencia.

Se utiliza cuando se no se cumple el supuesto de proporcionalidad de riesgos y se caracteriza por detectar diferencias entre las curvas de supervivencia cuando se cortan, aunque solamente al comienzo, por lo que no se recomienda su uso para estudios de largo plazo.

3.2.3. Test de Tarone-Ware

A partir de la función utilizada para definir el test anterior, se puede obtener el test de Tarone-Ware con los pesos equivalentes al tiempo observado menos el tiempo esperado para cada tiempo t_j, es decir, la raíz cuadrada del número de individuos en riesgo, $\sqrt{r_j}$.

$$tw = \sum_{j=1}^{r} \sqrt{r_j}(d_{1j} - e_{1j}) = \sum_{j=1}^{r} \sqrt{r_j}\left(d_{1j} - r_{1j}\frac{d_j}{r_j}\right)$$

La varianza del estadístico quedaría definida de la siguiente manera

$$\widehat{Var}(tw) = \sum_{j=1}^{r} \frac{r_{1j}r_{2j}d_j(r_j - d_j)}{r_j(r_j - 1)}$$

Por lo tanto el estadístico para el test de Tarone-Ware está dado por

$$TW = \frac{tw^2}{\widehat{Var}(tw)} \sim \chi_1^2$$

De la misma manera que el anterior, este test se utiliza en el caso de que no se cumpla el supuesto de proporcionalidad de riesgos.

4. Regresión de Cox para variables independientes del tiempo

Hasta el momento se han comparado grupos de individuos mediante estimaciones de las curvas de supervivencia y test que verificaban su similitud o disimilitud, pero no se llegaba a conocer el grado en el cual se diferenciaban unos grupos de otros. Este es el objetivo por el cual se aplica la regresión de Cox, la estimación de la influencia de las covariables sobre el riesgo.

Cuando se trabaja con modelos de regresión, la variable respuesta asume una distribución concreta, por lo general una distribución normal, con la que se valora el efecto de las variables predictoras. Sin embargo, en análisis de supervivencia no siempre se tienen unos datos que se aproximan a una distribución conocida. Además, estos datos se caracterizan por la presencia de censuras y, por lo tanto, el análisis no se puede realizar a través de regresión lineal múltiple o regresión logística. Por ello, se establece el método de regresión de Cox, que aunque no suele usarse para estimar la función de supervivencia, posibilita el conocimiento de la influencia de predictores en la variable respuesta y permite realizar estudios con datos de supervivencia que contienen observaciones censuradas.

Es una técnica multivariante capaz de reconocer y estimar la relación entre un conjunto de variables explicativas y la función de riesgo en el estudio y, simultáneamente, permite pronosticar las probabilidades de supervivencia partiendo de una distribución basal para un individuo en base a los valores que toman las variables explicativas.

El modelo de regresión de Cox se define a través de la función de riesgo, la cual depende del tiempo y de un conjunto de variables explicativas, variables independientes o covariables, $\boldsymbol{X} = (X_1, X_2, \dots, X_k)^T$, ya sean, cualitativas o cuantitativas, cuyos valores influyen en el tiempo hasta el evento, variable dependiente o respuesta.

El modelo de regresión de Cox (1972) está determinado por la siguiente función:

$$h(t|\boldsymbol{X}) = h_0(t) \cdot e^{(X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k)} = h_0(t) \cdot e^{\boldsymbol{X}^T\boldsymbol{\beta}}$$

Donde $h(t|\boldsymbol{X})$ es la función de riesgo del suceso a tiempo t para las diferentes covariables; $h_0(t)$ la función de riesgo basal a tiempo t; X_i las variables explicativas o predictoras y β_i los parámetros o coeficientes asociados a las covariables, con $i = 1, 2, \dots, k$. La función de riesgo basal es la función de riesgo para un individuo que toma todos los valores de las covariables iguales a cero. Este se designaría como individuo base o de referencia.

Este modelo puede ser estimado de forma semiparamétrica, en la que se tiene, por un lado, la función de riesgo basal, que solo depende del tiempo (parte no paramétrica), y por otro, la componente paramétrica, formada por una función exponencial de las covariables y sus parámetros, la cual determina el efecto de las mismas; o bien, de forma totalmente paramétrica, donde, el conjunto de la función de riesgo debe ser estimada mediante métodos paramétricos.

La estimación del riesgo basal puede realizarse mediante los métodos desarrollados durante el trabajo, ya sea, a partir del método producto límite de Kaplan-Meier, obteniendo así una estimación de la curva de supervivencia, o bien, de forma paramétrica, estableciendo una distribución que determine la función de riesgo.

Suponiendo una distribución exponencial de los datos y conociendo su función de riesgo, $h(t) = \lambda$, se puede obtener un modelo de regresión lineal en el que habría que estimar los parámetros β y tiene la siguiente forma:

$$h\,(t|\boldsymbol{X}) = \lambda \cdot e^{\boldsymbol{X}^T\boldsymbol{\beta}} \rightarrow \log h(t|\boldsymbol{X}) = \log\lambda + \boldsymbol{X}^T\boldsymbol{\beta}$$

En cuanto a la parte paramétrica, ésta solo depende de las covariables incluidas en el modelo, que se suponen independientes del tiempo durante todo el estudio. Lo que busca el modelo de regresión de Cox, es la estimación de los coeficientes β_i, con los que poder obtener información sobre la relación que existe entre dos grupos de individuos a comparar.

Suponiendo que se tienen dos valores de una covariable, X_i, el cociente de riesgos valorados en esos valores de la covariable, dejando fijos los valores de las demás covariables, viene determinado por:

$$HR = \frac{h(t|X_i = a)}{h(t|X_i = b)} = \frac{h_0(t)\cdot e^{a\boldsymbol{\beta}}}{h_0(t)\cdot e^{b\boldsymbol{\beta}}} = e^{\boldsymbol{\beta}(a-b)}$$

Como se puede comprobar, esta proporción no depende de la función de riesgo basal, es independiente del tiempo, y únicamente depende del valor de la covariable y su parámetro correspondiente. Por consiguiente, esta proporción de riesgos es una constante que no varía con el transcurso del tiempo, demostrando:

$$h(t|X_i = a) = cte \cdot h(t|X_i = b)$$

Lo que, implica que, una condición necesaria para que sea válido el modelo de Cox, es que los datos verifiquen la proporción entre los riesgos para la comparación de dos grupos. Por esto, el modelo de regresión de Cox también se denomina modelo de riesgos proporcionales, supuesto principal que debe cumplir un modelo.

El *Hazard ratio* da un factor de ponderación entre los riesgos, que mide cómo aumenta o disminuye el riesgo del evento en función de determinadas condiciones. Para el caso de variables continuas, lo que representa esta razón de riesgos es el incremento en una unidad, suponiendo que se tiene la variable "edad" por años; equivalentemente, suponiendo intervalos quinquenales de la "edad" la razón de riesgos da el incremento en cinco unidades de la covariable; mientras que, en el caso de variables de tipo cualitativo, mide la proporción de riesgo de un valor de la covariable sobre el otro.

Tomando un ejemplo básico de una variable binaria, supongamos "sexo", por un lado *hombres*, codificados con 1 y, por otro lado, *mujeres*, codificadas con 0; se calcula la proporción de riesgos de un grupo sobre otro, en este caso, se tendrán las mujeres como grupo de referencia, por lo que:

$$HR_{H,M} = \frac{h(t|H)}{h(t|M)} = e^{\beta(1-0)} = e^{\beta} \rightarrow h(t|H) = e^{\beta} \cdot h(t|M)$$

Como se puede comprobar, el riesgo de los hombres es e^{β} veces el de las mujeres. De la misma manera se trabaja con variables con más de dos categorías, tomando un grupo base y obteniendo la proporción de riesgos para cada uno de los grupos en función de él. Para un caso de querer comparar dos o más variables independientes el proceso a seguir es similar, tomando como grupo de referencia las categorías 0 de cada covariable y realizando combinaciones de las mismas hasta formar los grupos para después calcular el *Hazard ratio* de cada combinación con el grupo de referencia como base. Tras ello, se pueden realizar las comparaciones entre el resto de grupos una vez conocido los riesgos de cada grupo con el de referencia.

La interpretación de la razón de riesgos según su resultado es:

- Valores próximos a 1 indican que la variable independiente no implica un cambio en la tasa de riesgo.
- Valores inferiores a 1 implican una disminución del riesgo y un aumento de la probabilidad de supervivencia, correspondiéndose con coeficientes β negativos y factores de protección, suponiendo, aunque no se da en todos los casos, un evento de carácter negativo, como el fallecimiento.
- Valores superiores a 1 determinan un aumento en la velocidad del suceso del evento, y se corresponden con coeficientes β positivos y factores de riesgo, en el caso de eventos perjudiciales.

A la hora de estimar los parámetros del modelo de regresión de Cox el método más correcto es el denominado *Estrategia «hacia atrás»* o *Backward*, que consiste en la introducción de todas las covariables del estudio a analizar y, a continuación, ir suprimiendo aquellas que no sean significativas en el modelo, hasta obtener un conjunto de los mejores predictores de la variable respuesta, la supervivencia. Esta metodología es la que se tendrá en cuenta a la hora de realizar la parte práctica del trabajo.

Para conseguir el objetivo de estimar el vector de parámetros $\boldsymbol{\beta}$, Cox propuso un método de estimación que no dependía de la función de riesgo basal y permitía hacer inferencia sobre los parámetros, estimando la influencia de las covariables sobre el modelo. Este método maximiza la denominada función de verosimilitud parcial.

Supongamos que se tienen n individuos en un estudio, r tiempos de evento, sin empates y, por consiguiente, $n - r$ tiempos de censura. Sean $t_1, t_2, \dots, t_r$ los tiempos de evento de los individuos, I_j los intervalos de tiempo entre cada evento, $I_j = (t_{j-1}, t_j]$ y R_j el conjunto de individuos en riesgo en el intervalo I_j, para $j = 1, 2, \dots, r$; es decir, los individuos que permanecen aún en estudio a tiempo t_{j-1}. En estas condiciones,

$$P\big(al\ individuo\ j\ le\ ocurre\ el\ evento\ en\ \ t_j \mid en\ R_j\ ocurre\ un\ evento\ en\ t_j\big)$$

usando el Teorema de Bayes se deduce

$$\frac{P\left(al\ individuo\ j\ le\ ocurre\ el\ evento\ en\ t_j, en\ R_j\ ocurre\ un\ evento\ en\ t_j\ \right)}{P\left(en\ R_j\ ocurre\ un\ evento\ en\ t_j\right)} = \frac{P\left(al\ individuo\ j\ en\ R_j\ le\ ocurre\ el\ evento\ en\ t_j\right)}{P\left(en\ R_j\ ocurre\ un\ evento\ en\ t_j\right)}$$

Donde se verifica que

$$P\left(en\ R_j\ ocurre\ un\ evento\ en\ t_j\right) = \sum_{k\in R_j} P(al\ individuo\ k\ en\ R_j\ le\ ocurre\ el\ evento\ en\ t_j)$$

Considerando esta probabilidad en un intervalo de tiempo ínfimo $(t_j, t_j + \Delta t)$, dividiendo numerador y denominador por el incremento del tiempo, Δt, y tomando el límite del cociente cuando $\Delta t \to 0$, se obtiene la siguiente expresión dependiente de la función de riesgo:

$$\frac{h\left(t_j, \boldsymbol{X}_j\right)}{\sum_{k\in R_j} h(t_j, \boldsymbol{X}_k)} = \frac{h_0(t_j)\cdot e^{\boldsymbol{X}_j\boldsymbol{\beta}}}{h_0(t_j)\cdot \sum_{k\in R_j} e^{\boldsymbol{X}_k\boldsymbol{\beta}}} = \frac{e^{\boldsymbol{X}_j\boldsymbol{\beta}}}{\sum_{k\in R_j} e^{\boldsymbol{X}_k\boldsymbol{\beta}}}$$

Siendo $\boldsymbol{X}_i$ el vector de los valores de las covariables del individuo i-ésimo.

Finalmente, tomando el producto de cada una de las contribuciones de los r tiempos de evento, pues las ocurrencias del evento en dos individuos se consideran independientes, se define la función de verosimilitud parcial como

$$L(\beta) = \prod_{j=1}^{r} \frac{e^{X_i\boldsymbol{\beta}}}{\sum_{k\in R_j} e^{X_k\boldsymbol{\beta}}}$$

De tal manera que, la estimación de los coeficientes $\boldsymbol{\beta}$ se obtiene maximizando la función de verosimilitud parcial, o equivalentemente, maximizando el logaritmo de dicha función. Desafortunadamente, su aplicación solo es válida cuando no hay tiempos de empate entre los individuos, esto es, cuando no existen dos individuos que tengan el mismo tiempo de evento, por el contrario, se debería utilizar una aproximación de la función, como por ejemplo la de Breslow o la de Efron.

Continuando con el ejemplo de la base de datos "*btrial*" y aplicando una regresión de Cox en función de la respuesta inmunohistoquímica de los pacientes, se ha obtenido el siguiente resultado por pantalla:

```
> cox.zph(btrial.cox)
       chisq df   p
im     0.713  1 0.4
GLOBAL 0.713  1 0.4
```

```
coxph(formula = Surv(time, death) ~ im, data = btrial)

              coef exp(coef) se(coef)    z      p
imPositivo 1.6397    5.1536   0.7419 2.21 0.0271

Likelihood ratio test=5.05  on 1 df, p=0.02465
n= 15, number of events= 8
```

Previo a la interpretación del modelo, se debe comprobar el supuesto de proporcionalidad de riesgos para los valores de la covariable *im*. Mediante la función *cox.zph* aplicada al propio modelo, comprobamos que se cumple dicha condición, puesto que no se rechaza la hipótesis nula de proporcionalidad (p-valor > 0.05), tanto para la covariable como para un resultado global, que se consideran iguales, ya que el modelo solo se compone de una covariable. Una vez comprobado el supuesto, se pasan a analizar los resultados del modelo.

Se obtiene un valor del parámetro $\beta = 1.6397$, que, sabiendo que el grupo base es el de la respuesta inmunohistoquímica negativa, se puede concluir que el grupo de individuos positivos tienen un riesgo 5.15 veces superior que el grupo control, grupo negativo.

Por tanto, $h_P(t) = 5.1536 \cdot h_N(t)$, siendo $h_P(t)$ el riesgo de respuesta inmunohistoquímica positiva y $h_N(t)$ el riesgo de respuesta inmunohistoquímica negativa, luego el incremento porcentual del riesgo viene determinado por

$$\frac{h_P(t) - h_N(t)}{h_N(t)} \cdot 100 = \frac{5.1536 \cdot h_N(t) - h_N(t)}{h_N(t)} \cdot 100 = (5.1536 - 1) \cdot 100 = 415.36\%$$

Lo que nos indica que el riesgo aumenta en un 415% para el grupo de respuesta positiva frente al negativo.

Como se puede comprobar, estos resultados son estadísticamente significativos, ya que se obtiene un p-valor por debajo de 0.05. Mediante el programa estadístico R se puede obtener una gráfica con la que representar la diferencia entre los riesgos:

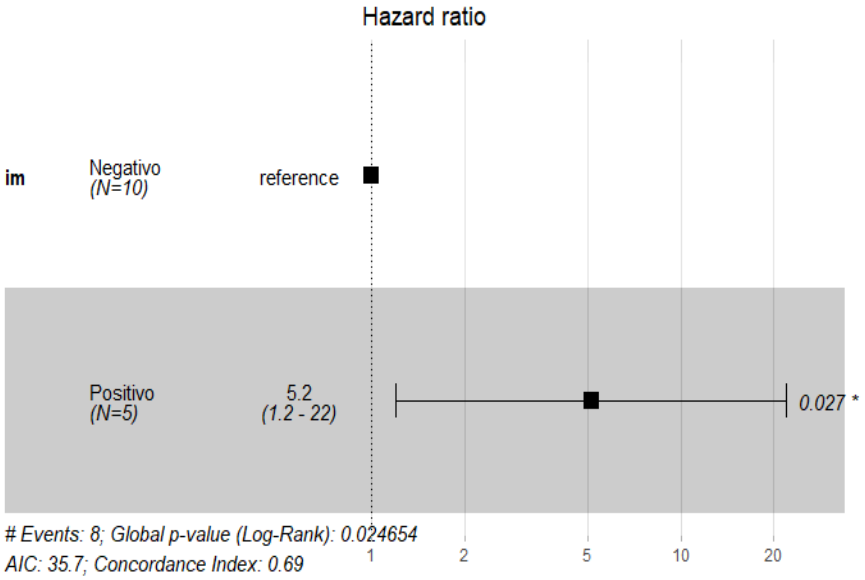


Gráfico 5. *Hazard ratio para la respuesta inmunohistoquímica positiva sobre negativa.*

5. Aplicación práctica

Finalmente, todo lo descrito en las secciones anteriores se va a poner en práctica mediante un ejemplo que abarca todos los contenidos hasta ahora desarrollados. La base de datos a utilizar se denomina *"bfeed"*, ésta pertenece al paquete *"KMsurv"* del programa estadístico R y es una de las muchas bases de datos que vienen implementadas en él.

La base de datos está formada por los tiempos para el destete de los recién nacidos que han sido alimentados con leche materna. En ella se tiene la información de 927 niños primogénitos nacidos después del año 1978 con una edad de gestación entre las 20 y 45 semanas, cuyas madres optaron por amamantarlos. La variable respuesta viene determinada por la duración de la lactancia materna en semanas, con un indicador que determina la lactancia materna completa o no, es decir, si el niño finalmente fue destetado.

Las variables de interés o explicativas son las siguientes:

- Raza: raza de la madre (1 si es blanca, 2 si es negra y 3 si es otra).
- Pobreza: indicador que determina si la madre se encuentra en situación de pobreza (1 si se encuentra en situación de pobreza y 0 en caso contrario).
- Tabaco: si la madre era fumadora al nacimiento del niño (1 si es afirmativa y 0 si es negativa).
- Alcohol: consumo de alcohol de la madre al nacer el bebé (1 si es cierta y 0 en sentido contrario).
- Edad de la madre: edad materna al nacimiento del niño (abarcando desde los 15 años hasta los 28).
- Año de nacimiento: año en el que nació el bebé (desde el año 1978 hasta el 1986, definidos con las últimas dos cifras).
- Años de escuela: número de años que la madre ha estado estudiando, medido en el nivel de educación (desde 3 años hasta 19 años).
- Atención prenatal después de 3 meses: necesidad de cuidado prenatal al tercer mes de embarazo (1 si la madre buscó atención prenatal y 0 si no la buscó).

A continuación se muestra un resumen de las variables, con el que se podrá tener una idea más general de los datos

```
      time             status           race      poverty     smoke      alcohol
 Min.   :  1.00   Min.   :0.0000   White:662   No :756   No :657   No :848
 1st Qu.:  4.00   1st Qu.:1.0000   Black:117   Yes:171   Yes:270   Yes: 79
 Median : 10.00   Median :1.0000   Other:148
 Mean   : 16.18   Mean   :0.9622
 3rd Qu.: 24.00   3rd Qu.:1.0000
 Max.   :192.00   Max.   :1.0000
```

```
     agemth          ybirth          yschool      pc3mth
 Min.   :15.00   Min.   :78.00   Min.   : 3.00   No :763
 1st Qu.:20.00   1st Qu.:80.00   1st Qu.:12.00   Yes:164
 Median :21.00   Median :82.00   Median :12.00
 Mean   :21.54   Mean   :81.97   Mean   :12.21
 3rd Qu.:23.00   3rd Qu.:84.00   3rd Qu.:13.00
 Max.   :28.00   Max.   :86.00   Max.   :19.00
```

Comenzaremos el estudio con un histograma de la variable tiempo con el que comprobar si los datos siguen una distribución normal o, por el contrario, se distribuyen aleatoriamente. En el gráfico siguiente se puede comprobar que a simple vista el tiempo de lactancia no se distribuye normalmente, siendo la línea roja la distribución normal y la negra la distribución de densidad de la variable de interés

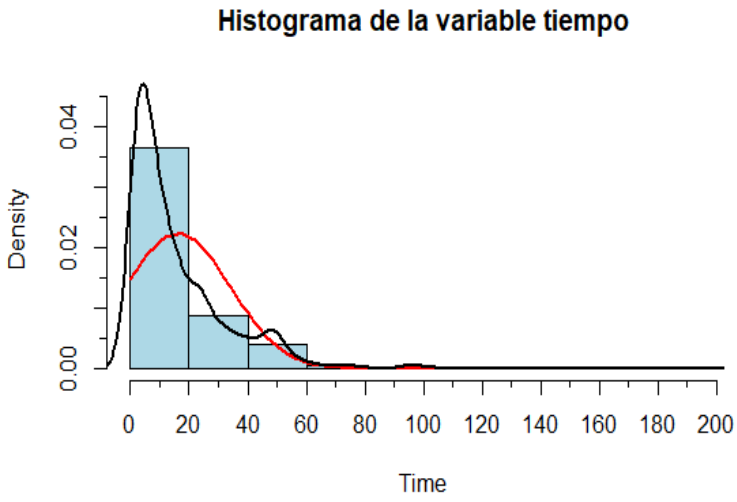


Gráfico 6. *Gráfico de densidad de la variable tiempo ("bfeed").*

Para verificar la normalidad o no de los datos, se realiza el test de Shapiro-Wilk, con el que se obtiene por pantalla un p-valor inferior a 0.05, por lo que se concluye el rechazo de la hipótesis nula, y con ello, el supuesto de normalidad, como se suponía en el gráfico.

El número de recién nacidos censurados, o bien, destetados, es de 35 niños, equivalente a un 3.78% de la muestra, siendo el resto de los 892 bebés no destetados (96.22%). Mediante el gráfico de censuras se puede observar una muestra aleatoria de 50 niños, en el que dos de ellos son censuras y el resto eventos. Aunque se muestre un máximo de aproximadamente 55 semanas, éstas ascienden hasta 192.

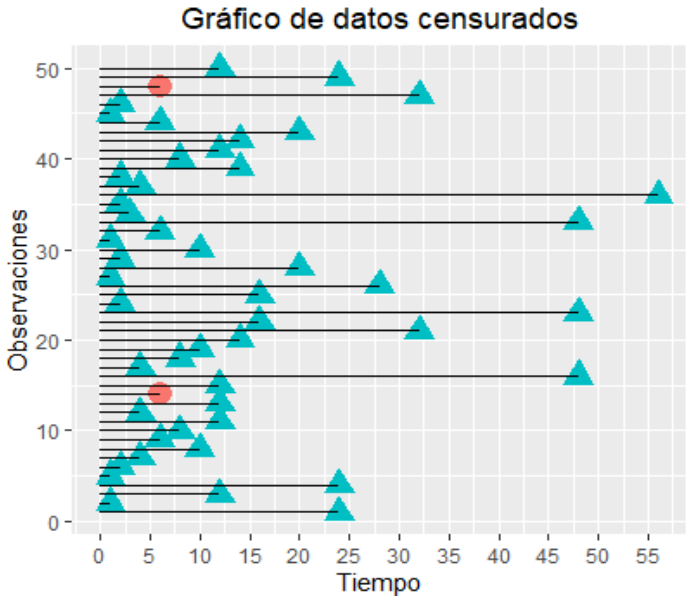


Gráfico 7. *Gráfico de datos censurados de la base de datos "bfeed".*

Comenzando con el análisis de la parte no paramétrica, la primera estimación que se debe realizar es la del estimador de Kaplan-Meier, con la que se obtienen las probabilidades de supervivencia para cada tiempo de evento. Como se tienen demasiados tiempos de evento, tan solo se mostrarán los 15 primeros para hacerse a la idea de la estimación de la función de supervivencia, que, posteriormente, se mostrará de manera gráfica (curva de supervivencia).

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	927	77	0.91694	0.00906	0.899342	0.93488
2	848	71	0.84016	0.01204	0.816889	0.86410
3	774	49	0.78698	0.01347	0.761020	0.81382
4	722	70	0.71068	0.01493	0.682003	0.74055
5	649	19	0.68987	0.01524	0.660640	0.72040
6	627	56	0.62826	0.01595	0.597763	0.66030
7	565	15	0.61158	0.01610	0.580829	0.64395
8	547	72	0.53108	0.01654	0.499631	0.56450
9	473	3	0.52771	0.01655	0.496252	0.56116
10	469	19	0.50633	0.01659	0.474840	0.53991
11	449	2	0.50407	0.01659	0.472584	0.53766
12	447	75	0.41950	0.01643	0.388497	0.45297
13	372	5	0.41386	0.01640	0.382927	0.44729
14	365	7	0.40592	0.01636	0.375090	0.43929
15	357	5	0.40024	0.01633	0.369481	0.43355

Observando la columna de supervivencia (*survival*), se tiene que la mediana se sitúa en la semana 12 del estudio, con 447 niños aún en riesgo y 75 eventos ocurridos al final de dicho intervalo de tiempo. La curva que determina la estimación de la función de supervivencia queda representada de la siguiente manera

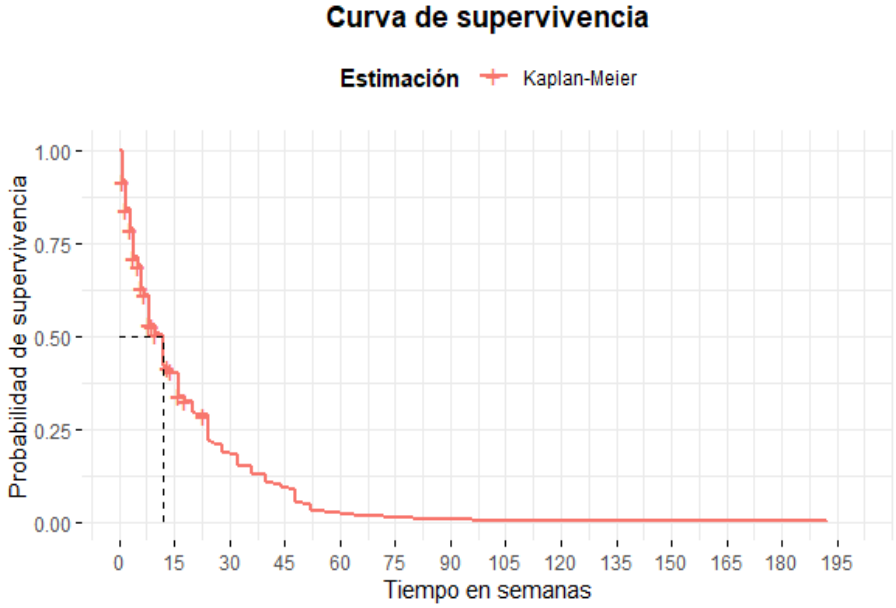


Gráfico 8. *Curva de supervivencia de la base de datos "bfeed".*

Como es trivial observar, la supervivencia de los niños va disminuyendo a medida que avanza el tiempo, ya que se van produciendo los destetes de los mismos, o bien, el abandono del estudio por causas desconocidas. Al comienzo del estudio se observa un decrecimiento pronunciado, haciendo ver la velocidad con la que se producen los eventos y censuras, éstas últimas marcadas con cruces en los tiempos correspondientes. A partir de la semana 45, dicho decrecimiento no es tan destacable, esto se debe al menor número de bebés en riesgo, ya que tan solo permanecen en el estudio 79 del total inicial. El estudio finaliza en la semana 192 con ningún recién nacido y, por consiguiente, con una supervivencia nula.

Si se observa el riesgo acumulado mediante la siguiente gráfica, se puede comprobar cómo va en aumento a lo largo del tiempo, destacando las primeras semanas, ya que en ellas se producen un mayor número de eventos entre los pequeños. También se pueden observar las curvas de supervivencia en porcentaje y en función del número de eventos acumulados, que se incluirán en el ANEXO (GRÁFICO 16. CURVA DE SUPERVIVENCIA EN PORCENTAJE ("BFEED"). y GRÁFICO 17. CURVA DE EVENTOS ACUMULADOS DE LA BASE DE DATOS "BFEED".**Gráfico 16**).

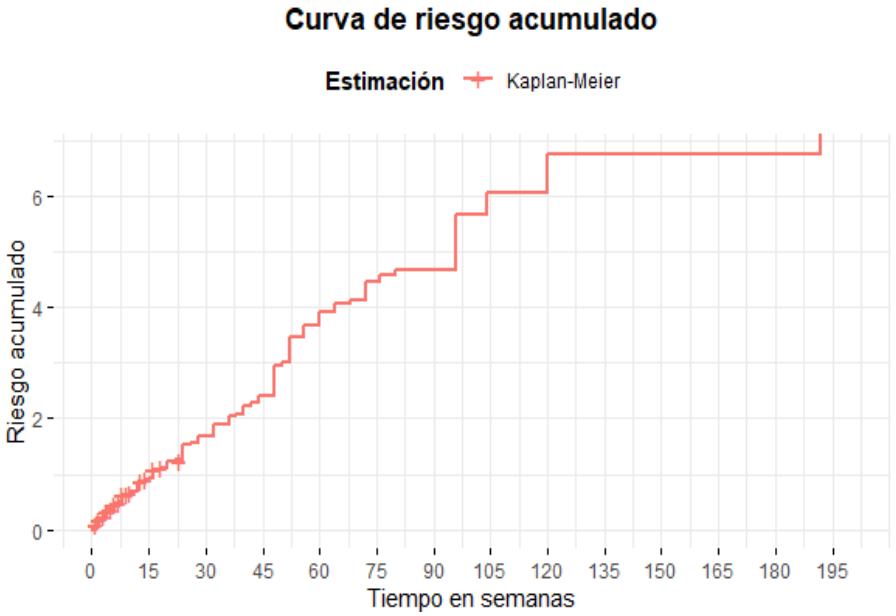


Gráfico 9. *Curva de riesgo acumulado ("bfeed").*

De la misma manera, se puede realizar la estimación de Kaplan-Meier con la influencia de covariables, obteniendo los resultados para las diferentes categorías de éstas. Por ejemplo, para calcular el estimador con la variable independiente *"race"*, los resultados son

```
> bfeed.km.race
Call: survfit(formula = Surv(time, status) ~ race, data = bfeed)

               n events median 0.95LCL 0.95UCL
race= White  662    634     12      10      12
race= Black  117    113      8       8      16
race= Other  148    145      8       6      12
```

Como se observa, la mediana para la raza blanca es mayor que para las otras categorías de la covariable, además que el número de recién nacidos es superior, lo que puede ser un indicio de diferencia entre las razas de las madres de los bebés. Realizando este cálculo para cada una de las covariables implicadas en el estudio, se puede comprobar que los valores de la mediana más destacados son 8 y 12, llegando a variar, en ocasiones, entre cifras de 4 a 22.

La suposición de que exista diferencia entre los valores de las variables explicativas se puede comprobar mediante pruebas no paramétricas realizadas a través de R, aplicando la función *survdiff* para las pruebas estadísticas Log-rank o de Wilcoxon, ésta última con el parámetro de la función *rho = 1*, o bien, la función *logrank_test* para las pruebas Log-rank, de Wilcoxon con el tipo de test a realizar equivalente al de "Gehan-Breslow" y la prueba de Tarone-Ware con dicho nombre como parámetro.

En la tabla 5, se resume la significación de los parámetros en función de las pruebas no paramétricas aplicadas a cada una de las diferentes covariables. La hipótesis nula de la que parten todos los test es la igualdad entre los grupos que forman la covariable, por el contrario, un p-valor inferior a 0.05 indica la significatividad del parámetro de dicha covariable y que, por lo tanto, existen diferencias entre los grupos que la componen.

Tabla 5. *Tabla de los test no paramétricos de comparación de dos o más grupos para las covariables de la base de datos "bfeed".*

	Test Log-rank	**Test de Wilcoxon**	**Test de Tarone-Ware**
race	**0.02**	**0.008**	**0.00887**
poverty	0.4	0.9	0.766
smoke	**0.001**	**0.004**	**0.0024**
alcohol	0.2	0.2	0.1742
agemth	0.2	0.2	x
ybirth	**0.007**	0.1	x
yschool	0.2	**0.04**	x
pc3mth	0.7	0.5	0.6202

Según sus resultados podemos concluir que las variables *race* y *smoke* son estadísticamente significativas para todos los test. En cuanto a las covariables *ybirth* e *yschool*, simplemente son significativas para uno de los tres test, la primera de ellas para el test Log-rank y la otra para el test de Wilcoxon. Mientras, el resto de variables, tienen un p-valor por encima de 0.05, por lo que no se puede rechazar la hipótesis nula de igualdad entre los grupos.

Si observamos gráficamente aquellas covariables con diferencias entre sus categorías

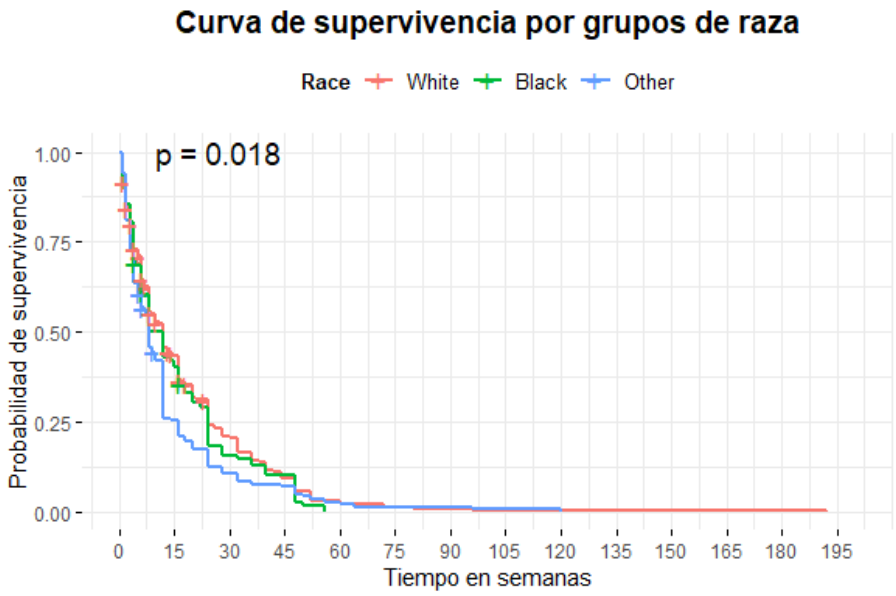


Gráfico 10. *Curva de supervivencia por grupos de raza ("bfeed").*

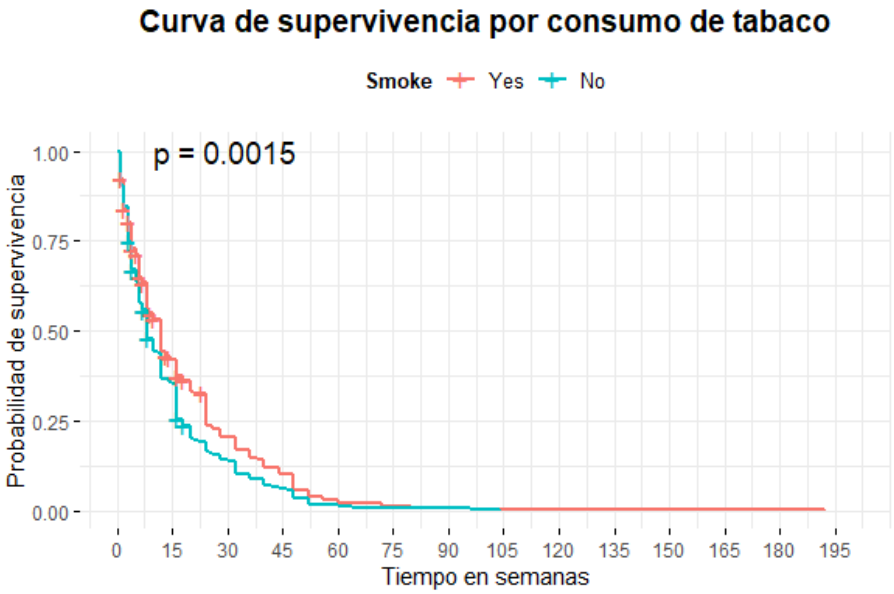


Gráfico 11. *Curva de supervivencia por consumo de tabaco ("bfeed").*

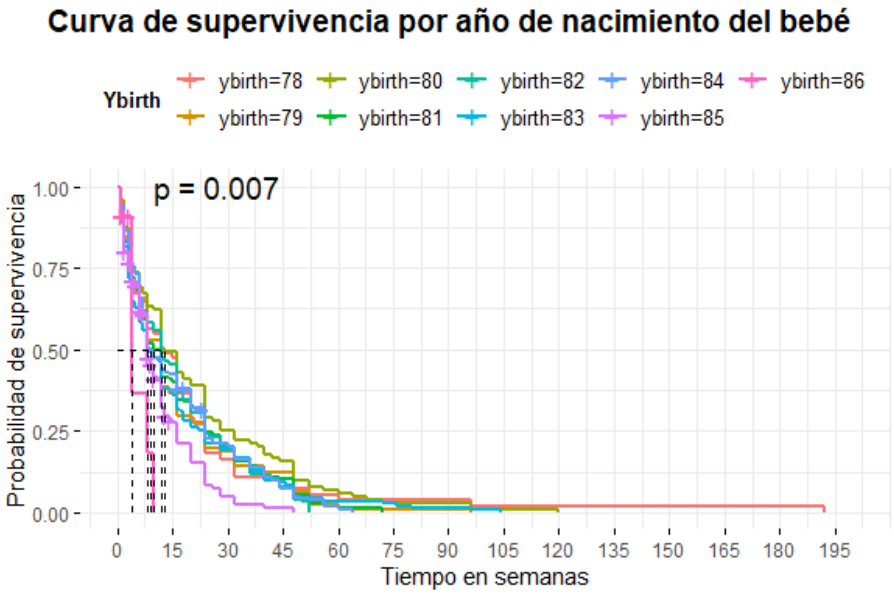


Gráfico 12. *Curva de supervivencia por año de nacimiento del bebé ("bfeed").*

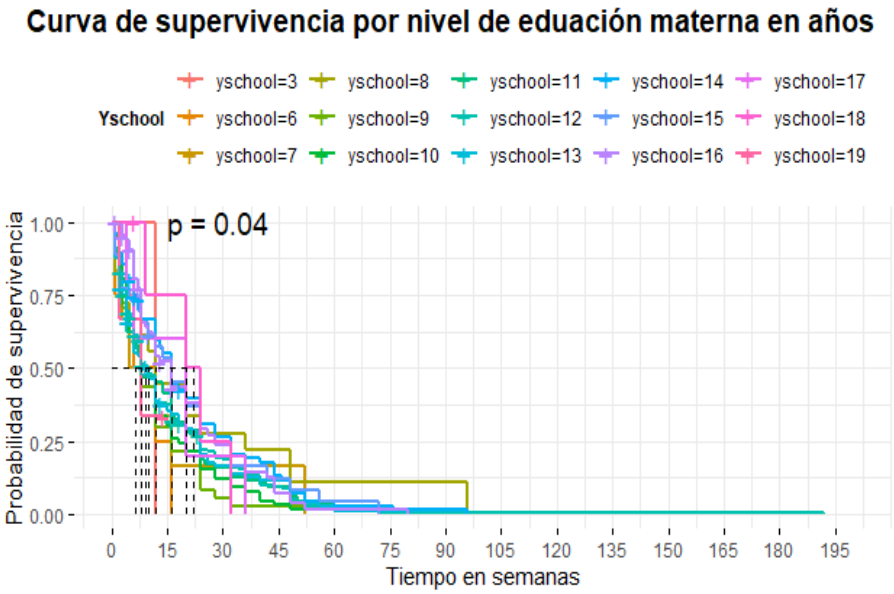


Gráfico 13. *Curva de supervivencia por nivel de educación materna en años ("bfeed").*

Pese a que gráficamente no se observan a simple vista las diferencias entre los grupos, las pruebas no paramétricas realizadas demuestran la disimilitud entre ellos, pero no se conoce el grado en el que difieren. Para ello, se va a realizar el último apartado de la práctica, lo denominado regresión de Cox.

Previo a la realización del modelo, se debe comprobar el supuesto de proporcionalidad para los valores de las variables independientes mediante la función *cox.zph*. Esta función se aplica para un modelo completo, es decir, un modelo con todas las covariables del estudio. El resultado por pantalla es el siguiente:

```
> cox.zph(bfeed.cox)
          chisq df      p
race     1.8973  2 0.3873
poverty  2.8255  1 0.0928
smoke    0.1554  1 0.6934
alcohol  0.0481  1 0.8263
agemth   3.7256  1 0.0536
ybirth   0.7681  1 0.3808
yschool 10.0658  1 0.0015
pc3mth   0.8394  1 0.3596
GLOBAL  11.4124  9 0.2485
```

La condición de riesgos proporcionales se cumple para todas las covariables, exceptuando aquella que determina los años de estudio de la madre. Para el resto de ellas, el p-valor es superior a 0.05 y se tiene que no hay evidencias suficientes para rechazar la hipótesis nula de proporcionalidad. Esto quiere decir que la variable significativa (*yschool*) debe ser introducida en el modelo como una covariable de estratificación, ya que permite corregir el modelo cuando alguna de las covariables no cumple el supuesto.

Comprobando el supuesto de proporcionalidad para el nuevo modelo se observa que, en este caso, todas las variables explicativas tienen un p-valor muy por encima de 0.05, por lo que la condición de proporcionalidad se cumple para el modelo y ya se puede realizar su interpretación.

```
Call:
coxph(formula = Surv(time, status) ~ race + poverty + smoke +
    alcohol + agemth + ybirth + strata(yschool) + pc3mth, data = bfeed)

                coef exp(coef) se(coef)      z        p
race Black   0.17253   1.18831  0.10724  1.609  0.10765
race Other   0.30278   1.35362  0.09989  3.031  0.00243
poverty Yes -0.14176   0.86783  0.09574 -1.481  0.13869
smoke Yes    0.22282   1.24959  0.08158  2.731  0.00631
alcohol Yes  0.19074   1.21014  0.12687  1.503  0.13272
agemth      -0.01266   0.98742  0.01945 -0.651  0.51506
ybirth       0.08150   1.08491  0.02081  3.917 8.97e-05
pc3mth Yes  -0.06799   0.93427  0.09309 -0.730  0.46515

Likelihood ratio test=37.67  on 8 df, p=8.653e-06
n= 927, number of events= 892
```

Como se observa en la última columna de los resultados obtenidos al realizar el modelo, tan solo tres covariables tienen un p-valor significativo, es decir, la raza materna, si la madre es fumadora o no y el año de nacimiento del bebé son las variables que están más relacionadas con la lactancia materna.

Como se mencionó anteriormente, la estimación de los parámetros del modelo se realizará a través de la *Estrategia «hacia atrás»* o *Backward*, en la que se irán eliminando las variables más significativas hasta obtener un modelo completamente significativo. Los pasos realizados se muestran en la siguiente tabla marcando aquellas variables que han sido suprimidas junto con su p-valor:

Tabla 6. *Método Backward de selección de variables para un modelo de regresión de Cox ("bfeed").*

Paso	**1**	**2**	**3**	**4**	**5**
race Black					✓
race Other					
poverty Yes				✓	
smoke Yes					
alcohol Yes			✓		
agemth	✓				
ybirth					
pc3mth Yes		✓			
p-valor	0.5151	0.4571	0.1397	0.1711	0.1319

Las covariables *race*, *smoke* e *ybirth* son las obtenidas para el modelo, es decir, son aquellas que mejor explican la lactancia materna para este conjunto de datos. Finalmente, pasamos a estimar el modelo y a interpretar sus resultados, ya que todas ellas son significativas y además, cumplen el supuesto de proporcionalidad.

```
Call:
coxph(formula = Surv(time, status) ~ I(race == " Other") + smoke +
    ybirth + strata(yschool), data = bfeed)

                            coef exp(coef) se(coef)     z        p
I(race == " Other")TRUE 0.26998   1.30994  0.09809 2.752  0.00592
smoke Yes               0.21372   1.23828  0.07938 2.692  0.00709
ybirth                  0.07412   1.07694  0.01817 4.080 4.51e-05

Likelihood ratio test=30.49  on 3 df, p=1.088e-06
n= 927, number of events= 892
```

El resultado principal a tener en cuenta es el p-valor global, obtenido a través de la prueba de razón de verosimilitud, con un valor muy por debajo de 0.05, por lo tanto significativo; con ello, se rechaza la hipótesis nula de que el vector de variables del modelo es cero (H_0: $\boldsymbol{\beta} = 0$) y se concluye que este modelo tiene sentido explicarlo mediante las covariables utilizadas. Cabe destacar que aquella covariable introducida como de estratificación, ya que sus categorías no eran proporcionales, no aparece en modelo y, por lo tanto, no tiene coeficiente con el que medir su efecto sobre la lactancia materna.

Pasando a realizar una interpretación de los coeficientes obtenidos por el modelo para las covariables significativas, se tiene:

- *Race (Other)*: tiene un coeficiente de regresión estimado de 0.26998, que, al ser positivo, indica un aumento en la tasa de riesgo cuando la covariable toma dicho valor (si la madre es de una raza distinta a negra o blanca, la probabilidad de una lactancia completa disminuye, es decir, el riesgo de ser destetado aumenta). El valor del riesgo de lactancia materna según la raza de la madre se comprueba mediante el exponencial del coeficiente, el cual tiene un valor de 1.3099. De este valor se concluye que el riesgo de destete cuando la raza de la madre es distinta de blanca o negra aumenta en 1.31 veces o en un 31% con respecto a la raza maternal de referencia, la blanca, manteniendo el resto de covariables constantes.

- *Smoke*: el coeficiente para esta covariable es de 0.2137, equivalente a un valor de *Hazard ratio* o exponencial de dicho valor de 1.2383, que indica que, manteniendo las demás covariables constantes, el riesgo de destete para los recién nacidos cuyas madres son fumadoras aumenta en un 1.24 o en un 24% con respecto a las que no fuman.

- *Ybirth*: su coeficiente de regresión es 0.0741, lo que indica un aumento en la tasa de riesgo cuando la covariable aumenta para cada año, es decir, un año de nacimiento más tardío del recién nacido implica una mayor probabilidad de destete. El *Hazard ratio* vale 1.0769, lo que se interpreta como que, por cada año que aumenta el año de nacimiento del bebé, el riesgo de destete aumenta en 1.08 veces o en un 7.7%, cuando el resto de covariables se mantienen constantes.

6. Conclusiones

El Trabajo de Fin de Grado desarrollado ha consistido en una introducción hacia el análisis de supervivencia, en el que se pretendía estudiar la variable principal que lo compone, "tiempo hasta un evento"; además de, la posible relación existente entre dicha variable respuesta y las covariables o variables independientes, con el propósito de buscar los modelos matemáticos más adecuados con los que conocer la influencia en el tiempo de supervivencia en función de dichas covariables, siempre no dependientes del tiempo. Para alcanzar este objetivo se han utilizado técnicas estadísticas propias del análisis de supervivencia, ya que los datos procedentes de este tipo de estudios destacan por la presencia de censuras o por sus funciones características de supervivencia o riesgo.

En primer lugar, de manera paramétrica, teniendo en cuenta la distribución que sigue la función de riesgo, por ejemplo, si tuviese una forma constante, seguiría una distribución Exponencial, como se vio anteriormente. Pero este método no es muy útil, ya que la función de riesgo puede llegar a tener una forma completamente aleatoria, a partir de la cuál no se puede llegar a obtener una distribución determinada, o bien, porque la hipótesis de que la función de riesgo siga una distribución concreta puede llegar a ser una selección errónea. Consecuentemente, se deciden aplicar métodos no paramétricos para evitar este tipo de problemas.

En cuanto a la aplicación de métodos no paramétricos, parten de la ventaja de no seguir una distribución determinada, por ello son estimaciones más utilizadas que las anteriores. En el presente trabajo, se ha destacado el método producto límite de Kaplan-Meier, un estimador bastante preciso con el que se calcula la probabilidad de supervivencia para cada instante de tiempo en el que se produce un evento, es decir, como ya se destacó, es un estimador de la función de supervivencia. Por su facilidad de cálculo y su precisión, ya que es un estimador máximo verosímil de la función de supervivencia, es el más utilizado en análisis de supervivencia. Además, en análisis no paramétrico se incluye la comparación de curvas de supervivencia entre dos o más grupos mediante pruebas estadísticas. Con ellas, simplemente se obtienen hipótesis de igualdad o diferencia de curvas, pero no el grado en el que difieren éstas. Por ello, este método no llega a ser el más destacado, pasando a combinar estimaciones paramétricas y no paramétricas con las que conocer la influencia de ciertas covariables sobre la variable dependiente.

Finalmente, se presentó un estimación semiparamétrica, la Regresión de Cox, formada por una función de riesgo basal (parte no paramétrica) y una función exponencial compuesta por las covariables y sus parámetros (parte paramétrica). A partir de ella, se podía llegar a conocer el coeficiente de ponderación que medía cuánto aumentaba o disminuía el riesgo de sufrir un evento en función de dichas covariables. La única condición que debe cumplir es el supuesto de proporcionalidad de riesgos para la comparación de dos grupos. Este método, con el que se calcula un modelo de regresión semiparamétrico, es el más señalado, ya que permite dar un valor con el que medir el nivel de riesgo de un grupo de individuos sobre otro.

En relación a los resultados obtenidos de la aplicación de los métodos sobre las bases de datos *btrial* y *bfeed*, se pueden concluir interpretaciones favorables, ya que se alcanzan modelos adecuados con los que explicar, en ambos casos, la variable respuesta del tiempo en función de las covariables en estudio. Para la primera de ellas, con un valor de significación del 0.0247, se concluye como significativo el modelo resultante con la inclusión de la covariable *im*, destacando el riesgo 5.15 veces superior para el grupo de respuesta inmunohistoquímica positiva con respecto a la respuesta negativa. Para la segunda base de datos, se incluyeron un mayor número de covariables, por lo que se suponía menos probable obtener un modelo completo significativo. Finalmente, tres de ellas resultaron ser las covariables que explicaban mejor la variable respuesta de lactancia materna completa (*race*, *smoke* e *ybirth*), con una significación muy por debajo de 0.05.

Bibliografía

Abraira, V. (2004). Análisis del tiempo hasta un evento (supervivencia). *Notas Estadísticas: Unidad de Bioestadística Clínica. Hospital Ramón y Cajal. Madrid.*, *30*(5), 223–225.

Análisis de supervivencia. (2017). 1–116.

Avedaño Garrido, M. L. (2013). *Modelos Generalizados de Riesgos Proporcionales para el análisis de supervivencia*. Universidad Complutense de Madrid.

Barrenechea López, L. (2018). *Técnicas no paramétricas y modelos de regresión para datos de tiempo de vida* (pp. 1–86).

Barrera Rebellon, M. (2008). *Análisis de supervivencia aplicado al problema de la deserción estudiantil en la universidad tecnológica de Pereira* (Issue July, p. 82).

Barroeta Rojo, C. (2016). *Modelos para el análisis de supervivencia en tiempos discretos: aplicación en el área de veterinaria*. Universidad de Barcelona.

Bellón, J. M. (2010a). *Análisis de supervivencia (I)*. EMEI. https://epidemiologiamolecular.com/analisis-supervivencia-i/

Bellón, J. M. (2010b). *Análisis de supervivencia (II)*. EMEI. https://epidemiologiamolecular.com/analisis-supervivencia-ii/

Boj del Val, E. (2017). *El modelo de regresión de Cox* (p. 49).

Brun González, L. P., & Salazar Uribe, J. C. (2016). Efecto de la censura informativa sobre la potencia de algunas pruebas tipo Log-Rank. *Ciencia En Desarrollo*, *7*(1), 45–53. https://doi.org/10.19053/01217488.4230

Carrasco, J. L. (n.d.). *El análisi estadístico de la supervivencia*.

Casas, B. Q. (2017). *Estadística en variables con censura: aplicación a datos medioambientales*. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/63786/6/bquintanillacTFM0617memoria.pdf

Castro-kuriss, C. (2018). *Análisis de Supervivencia mediante el empleo de R . Análisis de tiempos hasta un evento.* (Issue May) [Universidad de Buenos Aires]. https://doi.org/10.13140/RG.2.2.12736.84483/2

Cobo, E., González Alastrué, J. A., Muñoz García, P., Bigorra Llosas, J., Corchero García, C., Miras Rigol, F., Selva O´Callaghan, A., & Videla Ces, S. (2007). *Bioestadística para no estadísticos*.

Del Campo Esteban, R. (2015). *Diseño óptimo de Experimentos para el Análisis de Supervivencia* (p. 149).

Domènech, J. M. (1992). Una aplicación del análisis de la supervivencia en ciencias de la salud. *Anuario de Psicologia*, *55*, 109–142.

F. Lawless, J. (2003). Statistical Models and Methods for Lifetime Data. In *Angewandte Chemie International Edition, 6(11), 951–952.*

Fernandez, M., Abraira, V., Quereda, C., & Ortuno, J. (1996). Curvas de supervivencia y modelos de regresion: Errores y aciertos en la metodologia de aplicación. *Nefrologia*, *16*(5), 383–390.

Flores-Luna, L., & Salazar-Martínez, E. (2000). Análisis de supervivencia. Aplicación en una muestra de mujeres con cáncer cervical en México. *Salud Pública de México*, *42*. chrome-extension://dagcmkpagjlhakfdhnbomgmjdpkdklff/enhanced-reader.html?openApp&pdf=https%3A%2F%2Fwww.scielosp.org%2Farticle%2Fssm%2Fcontent%2Fraw%2F%3Fresource_ssm_path%3D%2Fmedia%2Fassets%2Fspm%2Fv42n3%2F2859.pdf

Flores-Luna, L., Zamora Muñoz, S., Slazar-Martínez, E., & Lazcano-Ponce, E. (2000). Aplicación en una muestra de mujeres con cáncer cervical en México. *Salud Pública de México*, *42*(3), 242–251.

Fuentelsaz, L., Gómez, J., & Polo, Y. (2004). Aplicaciones del análisis de supervivencia a la investigación en economía de la empresa. *Cuaderno de Economía y Dirección de La Empresa*, *19*, 81–114.

García Herrera, G. (2019). *Practica 3*. http://rstudio-pubs-static.s3.amazonaws.com/524797_db37a89e82ae4d0ab9047e87a939aacf.html

Godoy Aguilar, Á. M. (2009). *Introducción al Análisis de Supervivencia con R* [Universidad Nacional Autónoma de México]. http://www.cidpae.org.mx/documentos/documentos06.pdf

Gramatges Ortiz, A. (2002). Aplicación y técnicas del análisis de supervivencia en las investigaciones clínicas. *Revista Cubana de Hmatología, Inmunología y Hemoterapia*. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-02892002000200004

Hernández, L. (2020). *Análisis de supervivencia*. Doctor Mectrics by Metriplica. https://www.doctormetrics.com/analisis-de-supervivencia/

Herranz Valera, J. (2015). *Introducción al Análisis de Supervivencia con R* (pp. 12–15). http://www.cidpae.org.mx/documentos/documentos06.pdf

Janeiro, D., Portolés, J., Lopez-Sanchez, P., Tornero, F., Felipe, C., Castellano, I., Rivera, M., Fernandez-Cusicanqui, J., Cirugeda, A., Fernandez-Reyes, M. J., Rodriguez-Palomares, J. R., Bajo, M. A., Caparrós, G., & Ortiz, A. (2016). Cómo debemos analizar y describir la mortalidad de nuestros pacientes: experiencia del Grupo Centro Diálisis Peritoneal. *Nefrología*, *36*(2), 149–155. https://doi.org/10.1016/j.nefro.2015.09.014

Jiezhi Qi. (2009). Comparison of Proportional Hazards and Accelerated Failure Time Models College of Graduate Studies and Research in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Permission to Use. In *Master Thesis*. Saskatchewan.

José, B. S., Pérez, E., & Madero, R. (2009). Métodos estadísticos en estudios de supervivencia. *Anales de Pediatria Continuada*, *7*(1), 55–59. https://doi.org/10.1016/S1696-2818(09)70453-6

Karadeniz, P. G., & Ercan, I. (2017). Examining tests for comparing survival curves with right

censored data. *Statistics in Transition*, *18*(2), 311–328. https://doi.org/10.21307/stattrans-2016-072

Kassambara, A., Kosinski, M., & Biecek, P. (n.d.). *Forest Plot for Cox Proportional Hazards Model*. https://rpkgs.datanovia.com/survminer/reference/ggforest.html

López Vila, M. (2014). *Modelización matemática del riesgo de los accidentes de tráfico in-itinere para trabajadores de la Generalitat Valenciana*.

M. Molinero, L. (2004). Utilizando los modelos de supervivencia. *Www.Seh-Lelha.Org*, *Septiembre*, 1–6. d:%5CMarzo_2004%5CBibliografia%5CUtilizando los modelos de supervivencia.pdf

Martínez-González, M. Á., Alonso, Á., & Fidalgo, J. L. (2008). ¿Qué es un hazard ratio? Nociones de análisis de supervivencia. *Medicina Clinica*, *131*(2), 65–72. https://doi.org/10.1157/13123495

Martínez, J. (2017). *Análisis de Supervivencia en R*. http://rstudio-pubs-static.s3.amazonaws.com/316989_83cbe556125645b698c9ff6cf88c4c1a.html

Martorell, C., & do Pazo, F. (2018). *Análisis de supervivencia* (pp. 1–158).

Modelos de supervivencia. (n.d.). Modelos Estadísticos. Grado Biotecnología. Retrieved June 20, 2021, from https://rstudio-pubs-static.s3.amazonaws.com/375297_34390ade0ddb4dd2bbe3bf1abf884dfe.html#ejemplo_cáncer_de_pulmón

Molinero, L. M. (2001a). Modelos de regresión de Cox para el tiempo de supervivencia. *Asociación de La Sociedad Española de Hipertensión. Liga Española Para La Lucha Contra La Hipertensión Arterial.*, 1–4.

Molinero, L. M. (2001b). Tiempo hasta que ocurre un suceso. Análisis de supervivencia. *Asociación de La Sociedad Española de Hipertensión*, *June*, 1–8.

Molinero, L. M., & Ingeniería, A. (2004). Introducción al análisis de supervivencia cuando existen riesgos " competitivos " o cuando existen diferentes estados posibles (modelos multi − estado). *Asociación de La Sociedad Española de Hipertensión.*, 1–5.

Orbe, J., & Núñez-Antón, V. (2006). Alternative approaches to study lifetime data under different scenarios: From the PH to the modified semiparametric AFT model. *Computational Statistics and Data Analysis*, *50*(6), 1565–1582. https://doi.org/10.1016/j.csda.2005.01.010

Pita Fernández, S. (1995). Análisis de supervivencia. *Atención Primaria En La Red*, *2*(130–135), 1–14.

Practica 3. (n.d.). Retrieved June 20, 2021, from http://rstudio-pubs-static.s3.amazonaws.com/524797_db37a89e82ae4d0ab9047e87a939aacf.html

Pruenza García Hinojosa, C. (2014). *Estudio De Análisis De Supervivencia* (p. 88). https://repositorio.uam.es/bitstream/handle/10486/661556/pruenza_garcia_hinojosa_cristina_tfg.pdf?sequence=1

Quintanilla Casas, B. (2017). *Estadística en variables con censura: aplicación a datos medioambientales* [Univesitat Oberta de Catalunya]. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/63786/6/bquintanillacTFM0617memoria.pdf

Ramalle-Gómara, E. (2000). Modelos estadíticos para el análisis de la supervivencia. *Diálisis y Trasplante: Publicación Oficial de La Sociedad Española de Diálisis y Trasplante.*, *21*(1), 1–3. https://openlibra.com/es/book/download/estadistica-descriptiva-univariante

Rebasa, P. (2005). Conceptos básicos del análisis de supervivencia. *Cirugia Española*, *78*(4), 222–230. http://mail.aecirujanos.es/revisiones_cirugia/2005/Octubre2_2005.pdf

Rial Boubeta, A., & Varela Mallou, J. (2008). *Estadistica práctica para la investigación en ciencias de la salud*.

Ríos Vargas, J. Á. (2017). *Análisis de supervivencia* (p. 31).

Rodríguez Jaume, M. J., & Mora Catalá, R. (2001). Capítulo 12. Análisis de supervivencia. In *La muerte y la máscara en Pablo Picasso* (pp. 213–226). https://doi.org/10.3726/978-1-4539-1172-3/14

Salazar Uribe, J. C., García Cruz, E. K., Gaviria Peña, C., & Guarín Escudero, V. (2020). Introducción al análisis de supervivencia avanzada. In *Introducción al análisis de supervivencia avanzada*. https://doi.org/10.21500/9789588474939

Salinas F., M. (2008). Modelos de regresión VI. Análisis de Supervivencia. *Ciencia & Trabajo*, *October*, 75–78.

Subirana, I. (2020). *Análisis de datos longitudinales*. https://bookdown.org/isubirana/longitudinal_data_analyses/

Valencia-Orozco, A., Parra-Lara, L. G., Martínez, J. W., & Tovar-Cuevas, J. R. (2019). Aplicacion de modelos paramétricos alternativos para el análisis de supervivencia de pacientes con cáncer. *Revista Peru Medicina Salud Publica*, *36*(2), 341–348.

W. Hosmer, D., & Lemeshow, S. (1999). *Applied Survival Analysis. Regression Modeling of Time to Event Data*.

Yan, J. (2012). *Package "KMsurv". Data sets Klein and Moeschberger (1997), Survival Analysis*.

Zapata Acevedo, S. A. (2018). *Análisis estadístico de eventos asociados a variables de tiempo en R : Modelo de supervivencia en pacientes con carcinoma de células renales*. Universidad de Barcelona.

Summary

Survival analysis is the name given to the set of techniques that allow us to study the variable time until an event occurs and its dependence on other possible explanatory variables, considering the partial information contained in the censored observations.

The time to the event is the relevant variable in survival analysis, which represents the follow-up time of everyone, a term defined as the time elapsed from the time an individual enters the study until the occurrence of the event, or until the end of the study time or abandonment time. It should be noted that the inclusion time of everyone is usually different from the time of any other individual, because of staggered inclusion. The dates of everyone's last observation are determined by different causes, either because of the event or because of censoring. Censoring occurs in the study when it is not possible to determine in a precise way the time of occurrence of the fixed event, having then an incomplete information of certain individuals. They fall into three relevant categories:

- Right censoring: occurs when the time to event T is greater than in observation time or C censoring, that is, when the event has not occurred during the observation period. Within this set, three types of censoring stand out:

  Type I censoring. A specific duration of the study is defined and will be determined as censoring when the time of the event is greater than the end-of-study time.

  Type II censoring. A specific number of events that marks determined the end of the study, that is, if there are n observations, when the r-th event occurs ($r < n$), the study will be concluded. On some occasions, a final time is set when the desired proportion is not reached in a long period of time.

  Type III censoring. It presents with a random event during the study, it can occur at the time when information of the individual is lost, when an event different from the one predetermined occurs or when, even at the end of the study, the event has not occurred. It is also called non-informative or random.

- Left censoring: it occurs when the event occurs prior to the incorporation in the study, that is, when an individual is going to be analyzed for the first time and the event with which the study begins has already occurred and, in addition, its time of occurrence is unknown. Left censoring is characterized by the search for a second instant in which the event occurs, considering that the previous case happened out the study.

- Interval censoring: the time at which it occurs is unknown, we simply have the information that it occurs between two instants of time in an observation period. Interval censoring usually occurs in studies in which the dependent variable is divided into time intervals, so the event can occur within the interval, but the exact time is unknown.

Let T be a non-negative random variable denoting the survival time or time until the event occurs. The most relevant functions in survival analysis are:

Probability density function:

- Discrete case:

$$f(t_j) = P(T = t_j)\ with\ j = 1, 2 \ldots t_1 < t_2 < t_3 < \ \ldots and \sum_{t_j \in T} f(t_j) = 1\ and\ f(t_j) \geq 0$$

- Continuous case:

$$f(x)\colon \mathbb{R} \to \mathbb{R}\ such\ that\ P(X \in (a, b)) = \int_a^b f(x)dx\ and \int_\infty^\infty f(x)dx = 1\ with\ f(x) \geq 0$$

Distribution function:

- Discrete case: $F(t) = P(T \leq t) = \sum_{t_j \leq t} f(t_j)$
- Continuous case: $F(t) = P(T \leq t) = \int_0^t f(x)dx$

Survival function:

The survival probability is defined as the probability of remaining in the study without the event occurring on the individuals. Therefore, the survival function, or cumulative rate, is the probability that an individual in the study does not present the event before a given time; that is, that it survives until a specific time t.

Let T be a positive random variable with a distribution function $F(t)$ and probability density function $f(t)$; the survival function, $S(t)$, is defined by

$$S(t)\ =\ 1 - F(t)\ = P(T > t)$$

This function associates to each time t the probability that a subject survives at that instant of time. It should be noted that the function at $S(0) = 1$, that is, that an individual is always alive at the beginning of the study, while for an infinite time, such probability is zero; that is, $S(t) = 0$ when $t \to \infty$. Therefore, as the distribution function is increasing towards 1, it's observed that the survival function is a decreasing function. For a discrete random variable, one has that the survival function is the summation of all times greater than t of the density function; whereas for a continuous variable it's the integral for a time greater than t.

Risk function:

The hazard function represents the evolution of the probability of event in relation to the survival time of the individuals. In general, it's the probability that an individual who is being observed at time t the event will happen to him at that time, therefore, this function is the most appropriate for the description of the study, because the values constitute the incidence rates of the event analyzed.

- Discrete case: $h(t_j) = P(T = t_j \mid T \geq t_j) = \frac{P(T=t_j)}{P(T \geq t_j)} = \frac{f(t_j)}{S(t_j)}$
- Continuous case: $h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T \leq t+\Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$

The hazard function can be increasing, decreasing, constant or bathtub-shaped. Each of these forms is represented by different distributions.

Cumulative hazard function:

- Discrete case: $H(t) = \sum_{t_j \leq t} h(t_j)$
- Continuous case: $H(t) = \int_0^t h(x)\, dx$

The useful distributions that stand out in survival analysis are the Exponential and the Weibull, although this does not mean that they are always the best distributions because it is necessary to carry out a previous study to obtain the model that best fits the data. The shape of the hazard function will be the one that will give us preference for one model or another.

Assuming an Exponential distribution, $\varepsilon(\lambda)$, of positive parameter λ, its hazard function is determined by $h(t) = \lambda$, which indicates that the event risk does not change over time and, therefore, the property of *loss or absence of memory* is satisfied, not depending on the probability of event at time t of any previous time. In the case of a Weibull distribution, it is known that the hazard function varies over time and can have a random shape, depending on the values taken by its parameters of shape γ and scale λ. It is determined by $h(t) = \gamma\lambda t^{\gamma-1}$, where if $\gamma > 1$, it will be an increasing function, if $\gamma < 1$ it will be decreasing and if $\gamma = 1$ it will match with the hazard function of the Exponential model.

Due to the presence of censoring, it may not be appropriate to use the mean as an estimator of the value of the variable, being advisable to use the median or percentiles. However, other methods are needed to estimate survival, such as a nonparametric method that does not require a specific distribution. The most prominent method is the Kaplan-Meier limit product, which consists of an estimation of the survival function.

Marking the time elapsed between each event by an interval I_j, which runs from a time t_{j-1} to the event time t_j, $I_j = (t_{j-1}, t_j]$. For each event time t_j, we define, on the one hand, as r_j, the number of individuals, such that, meet $T > t_{j-1}$, that is, those that arrive alive at the interval I_j; and, on the other hand, d_j as the number of individuals that experience the event in that interval. Using the compound probability theorem (Product rule) we have

$$P(T > t_j) = P(T > t_{j-1}) \cdot P(T > t_j | T > t_{j-1}) \rightarrow S(t_j) = S(t_{j-1}) \cdot \frac{r_j - d_j}{r_j}$$

The conditional probability of surviving to a time t_j having survived to a previous time is called survival probability and is determined by r_j and d_j. Successively applying this probability for each time interval deduces the Kaplan-Meier estimator

$$Survival\ probability = \frac{r_j - d_j}{r_j} = 1 - \frac{d_j}{r_j} \quad \rightarrow \quad \hat{S}(t) = \prod_{t_j \leq t} \frac{r_j - d_j}{r_j}$$

It's a maximum likelihood estimator of the survival function and meets the properties of being unbiased, consistent and efficient of an estimator. From these results, the graphical representation obtained is an estimate of the survival function and is called the survival curve. It is a decreasing function, and its shape reveals the speed with which events occur as time evolves.

Comparisons can be made between two survival curves and, although at first glance differences between groups may be observed, they may not be true, therefore, it's appropriate to resort to nonparametric tests to contrast the equality of survival functions. The most notable nonparametric test is the Log-rank test, which consists of comparing the number of events observed in each group against the number of expected events, assuming that there are no differences between them, that is, considering H_0 true. The resulting statistic of

$$LR = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1} \ where\ k\ is\ the\ numer\ of\ groups$$

is compared with the Chi-square distribution and according to the result of the p-value, the existence of a statistical difference between the curves or not is determined. Another test, also used, is the generalized Wilcoxon test or Breslow's test, which is based on the weighted sum of the observed and expected differences between the number of events. The weights used are equivalent to the number of individuals at risk for each time t_j, that is, r_j. Assuming two groups, the expected number for the first group at time t_j would be $e_{1j} = r_{1j}d_j/r_j$ and the test statistic would be defined as follows

$$w = \sum_{j=1}^{r} r_j(d_{1j} - e_{1j}) = \sum_{j=1}^{r} r_j\left(d_{1j} - r_{1j}\frac{d_j}{r_j}\right); \widehat{Var}(w) = \sum_{j=1}^{r} \frac{r_{1j}r_{2j}d_j(r_j - d_j)}{r_j^2(r_j - 1)}$$

Therefore, the generalized Wilcoxon or Breslow's test statistic is given by

$$W = \frac{w^2}{\widehat{Var}(w)} \sim \chi^2_1$$

As a last nonparametric test, the Tarone-Ware test is developed, with the weights equivalent to the observed time minus the expected time for each time t_j, that is, the square root of the number of individuals at risk, $\sqrt{r_j}$. The statistic and its variance are determined by

$$tw = \sum_{j=1}^{r} \sqrt{r_j}(d_{1j} - e_{1j}) = \sum_{j=1}^{r} \sqrt{r_j}\left(d_{1j} - r_{1j}\frac{d_j}{r_j}\right);\ \widehat{Var}(tw) = \sum_{j=1}^{r} \frac{r_{1j}r_{2j}d_j(r_j - d_j)}{r_j(r_j - 1)}$$

So, the Tarone-Ware test statistic is defined by: $TW = \frac{tw^2}{\widehat{Var}(tw)} \sim \chi^2_1$

These last two tests are used in case the assumption of risk proportionality is not met.

Concluding with the theoretical part, the Cox regression for time-independent variables is presented, with the aim of finding out the degree to which some groups differ from others, that is, to obtain an estimate of the influence of the covariates on risk. This semiparametric model is defined through the risk function, which depends on time and a set of independent variables or covariates $\boldsymbol{X} = (X_1, X_2, \dots, X_k)^T$, whose values influence the time to the event. The Cox regression model (1972) is determined by the following function:

$$h(t|\boldsymbol{X}) = h_0(t) \cdot e^{(X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k)} = h_0(t) \cdot e^{\boldsymbol{X}^T\boldsymbol{\beta}}$$

Where $h(t|\boldsymbol{X})$ is the hazard function of the event at time t for the different covariates; $h_0(t)$ the baseline hazard function at time t; X_i the explanatory or predictor variables and β_i the parameters or coefficients associated with the covariates, with $i = 1, 2, \dots, k$. The baseline hazard function is the hazard function for an individual who takes all the values of the covariates equal to zero. This would be designated as the base or reference individual.

The estimation of the baseline hazard can be carried out using the methods developed during the work, either from the Kaplan-Meier limit product method, thus obtaining an estimate of the survival curve, or parametrically, by establishing a distribution that determines the hazard function. As for the parametric part, this only depends on the covariates included in the model, which are assumed to be independent of time throughout the study. What the Cox regression model seeks is the estimation of the coefficients β_i, with which to obtain information about the relationship between two groups of individuals to be compared.

Assuming that there are two values of a covariate, X_i, the ratio of risks valued at those values of the covariate, leaving the values of the other covariates fixed, is determined by:

$$HR = \frac{h(t|X_i = a)}{h(t|X_i = b)} = \frac{h_0(t) \cdot e^{a\boldsymbol{\beta}}}{h_0(t) \cdot e^{b\boldsymbol{\beta}}} = e^{\boldsymbol{\beta}(a-b)}$$

This weighting factor between risks is called *Hazard ratio* and measures the increase or decrease in the risk of the event based on certain conditions.

As can be seen, this proportion does not depend on the baseline hazard function (independent of time) and only depends on the value of the covariate and its corresponding parameter. Therefore, this proportion of risks is a constant that does not vary over time, showing: $h(t|X_i = a) = cte \cdot h(t|X_i = b)$. Which implies that, a necessary condition for the Cox model to be valid is that the data must verify the ratio between the risks for the comparison of two groups. For this reason, the Cox regression model is also called proportional hazards model, the main assumption that a model must meet.

The interpretation of the hazard ratio according to its outcome is:

- Values close to 1 indicate that the independent variable does not imply a change in the hazard ratio.
- Values lower than 1 imply a decrease in risk and an increase in the probability of survival, corresponding to negative β coefficients and protective factors, assuming, although this does not occur in all cases, an event of a negative nature, such as death.

- Values greater than 1 determine an increase in the speed of occurrence of the event and correspond to positive β coefficients and risk factors, in the case of detrimental events.

To estimate the parameters of the model, the *"Backward" strategy* will be used, which consists of introducing all the covariates of the study to be analyzed and then delete those variables that are not significant in the model, until a set of the best predictors of the response variable, survival, is obtained.

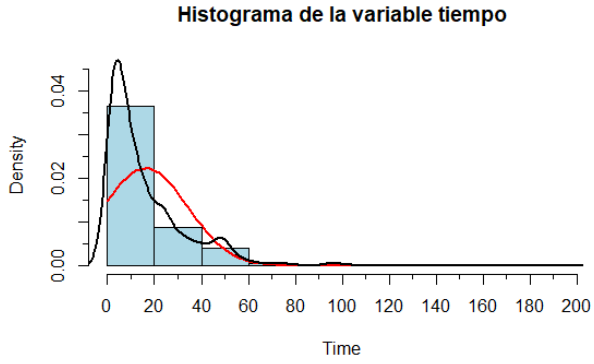
Cox proposed an estimation method that did not depend on the baseline hazard function and allowed inference to be made on the parameters, estimating the influence of the covariables on the model. This method maximizes the so-called partial likelihood function, defined as

$$L(\beta) = \prod_{j=1}^{r} \frac{e^{X_i \boldsymbol{\beta}}}{\sum_{k \in R_j} e^{X_k \boldsymbol{\beta}}}$$

Thus, the estimation of the coefficients β is obtained by maximizing the partial likelihood function, or equivalently, by maximizing the logarithm of this function.

Finally, all the methods described above will be presented in a practical way through a database called *bfeed*, obtained from the statistical program R. It is made up by times for the weaning of newborns who have been breastfed. It contains information on 927 first-born children whose mothers chose to breastfeed them. The response variable is determined by the duration of breastfeeding in weeks, with an indicator that determines complete breastfeeding or not, that is, whether the child was finally weaned. The explanatory variables are:

- Race: race of the mother (1 if white, 2 if black and 3 if other).
- Poverty: indicator of whether the mother is in poverty (1 in poverty and 0 otherwise).
- Smoke: whether the mother was a smoker at the birth of the child (1 if yes and 0 if no).
- Alcohol: mother's alcohol consumption at the birth of the child (1 if true and 0 otherwise).
- Mother's age: maternal age at child's birth (ranging from 15 to 28 years).
- Year of birth: year in which the baby was born (from 1978 to 1986).
- Years of schooling: years that the mother has been studying (from 3 years to 19 years).
- Prenatal care after 3 months: need for prenatal care at the third month of pregnancy (1 if the mother sought prenatal care and 0 if she did not).

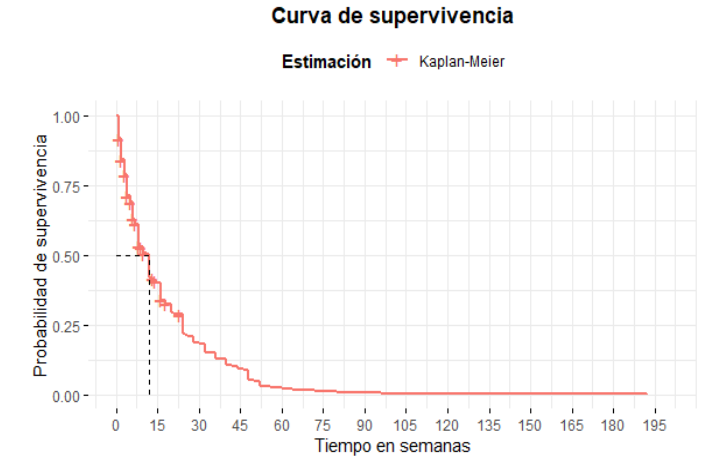


As can be seen immediately in the graph, the distribution of the variable lactation time is not normal. But applying the Shapiro-Wilk test, to demonstrate the non-normality of the data, we obtain a significance lower than 0.05, therefore, it is concluded that the data do not follow a normal distribution.

The number of censored newborns is 35, while the rest, 892 babies (96.22%), completed breastfeeding. The number of weeks under study is 192.

Applying the Kaplan-Meier estimation we obtain the following result per screen for the first 12 event times, together with the corresponding survival probabilities, which will allow us to make the survival curve graph.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	927	77	0.91694	0.00906	0.899342	0.93488
2	848	71	0.84016	0.01204	0.816889	0.86410
3	774	49	0.78698	0.01347	0.761020	0.81382
4	722	70	0.71068	0.01493	0.682003	0.74055
5	649	19	0.68987	0.01524	0.660640	0.72040
6	627	56	0.62826	0.01595	0.597763	0.66030
7	565	15	0.61158	0.01610	0.580829	0.64395
8	547	72	0.53108	0.01654	0.499631	0.56450
9	473	3	0.52771	0.01655	0.496252	0.56116
10	469	19	0.50633	0.01659	0.474840	0.53991
11	449	2	0.50407	0.01659	0.472584	0.53766
12	447	75	0.41950	0.01643	0.388497	0.45297



The median event time is at week 12, with 447 children at risk and 75 events occurring at the end of the interval. The curve starts with a steep decrease, as numerous events occur and, from week 45 onwards, the decrease is maintained but is not so remarkable, as the number of babies at risk has decreased to 79. In the same way, the Kaplan-Meier estimator can be applied with the influence of some covariate, obtaining the same results depending on the different covariate categories. With these results, we can assume differences between the groups, but to be sure, we applied the Log-rank, Wilcoxon and Tarone-Ware curve comparison tests to demonstrate the equality or difference between the curves.

The significance of the coefficients of the covariates as a function of the nonparametric tests mentioned above is summarized below:

	Test Log-rank	Test de Wilcoxon	Test de Tarone-Ware
race	**0.02**	**0.008**	**0.00887**
poverty	0.4	0.9	0.766
smoke	**0.001**	**0.004**	**0.0024**
alcohol	0.2	0.2	0.1742
agemth	0.2	0.2	x
ybirth	**0.007**	0.1	x
yschool	0.2	**0.04**	x
pc3mth	0.7	0.5	0.6202

With this, we can conclude that the variables *race*, *smoke*, *ybirth* and *yschool* are significant for one or more tests, therefore, covariates with differences between their groups are assumed. Finally, by applying Cox regression we can obtain the degree to which the groups differ for each covariate.

Before the model is run, the assumption of proportionality for the values of the independent variables must be checked. A significance below 0.05 is obtained for the variable yschool, therefore it cannot be used and must be introduced in the model as a stratification variable. Checking the proportionality assumption with the new model, the expected results are obtained, all covariates have a p-value above 0.05 and, therefore, meet the assumption. Next, independent variables with p-values above 0.05 will be eliminated, since what we are looking for is the significance of the coefficients of the covariates to obtain those that best explain the response variable. The final model obtained consists of the variables race, smoke and ybirth, since they are the only ones that maintain their significance.

```
Call:
coxph(formula = Surv(time, status) ~ I(race == " Other") + smoke +
    ybirth + strata(yschool), data = bfeed)

                              coef exp(coef) se(coef)     z        p
I(race == " Other")TRUE 0.26998   1.30994  0.09809 2.752  0.00592
smoke Yes               0.21372   1.23828  0.07938 2.692  0.00709
ybirth                  0.07412   1.07694  0.01817 4.080 4.51e-05

Likelihood ratio test=30.49  on 3 df, p=1.088e-06
n= 927, number of events= 892
```

The likelihood ratio test has a p-value lower than 0.05, therefore significant; with this, the null hypothesis that the vector of variables of the model is zero is rejected $(H_0: \boldsymbol{\beta} = 0)$ and it's concluded that this model makes sense to explain it by means of the covariates used. It should be noted that the covariate introduced as a stratification variable, since its categories were not proportional, does not appear in the model and, therefore, has no coefficient with which to measure its effect on breastfeeding.

The interpretation of their coefficients is as follows:

- *Race (Other):* it has an estimated regression coefficient of 0.26998, which indicates an increase in the risk rate when the covariate takes that value. The value of the risk of breastfeeding according to the mother's race is tested by the exponential of the coefficient, which has a value of 1.3099. From this value it is concluded that the risk of weaning when the mother's race is other than white or black increases by 1.31 times or 31% with respect to the reference maternal race, white, holding all other covariates constant.

- *Smoke:* the coefficient for this covariate is 0.2137, equivalent to a Hazard ratio or exponential value of that value of 1.2383, indicating that, holding all other covariates constant, the risk of weaning for newborns whose mothers are smokers increases by 1.24 or by 24% with respect to those who do not smoke.

- *Ybirth:* its regression coefficient is 0.0741, indicating an increase in hazard ratio when the covariate increases for each year, i.e., a later birth year of the newborn implies a higher probability of weaning. The Hazard ratio is 1.0769, which is interpreted to mean that, for each year that the year of birth of the infant increases, the risk of weaning increases by 1.08 times or 7.7%, when all other covariates are held constant.

ANEXO

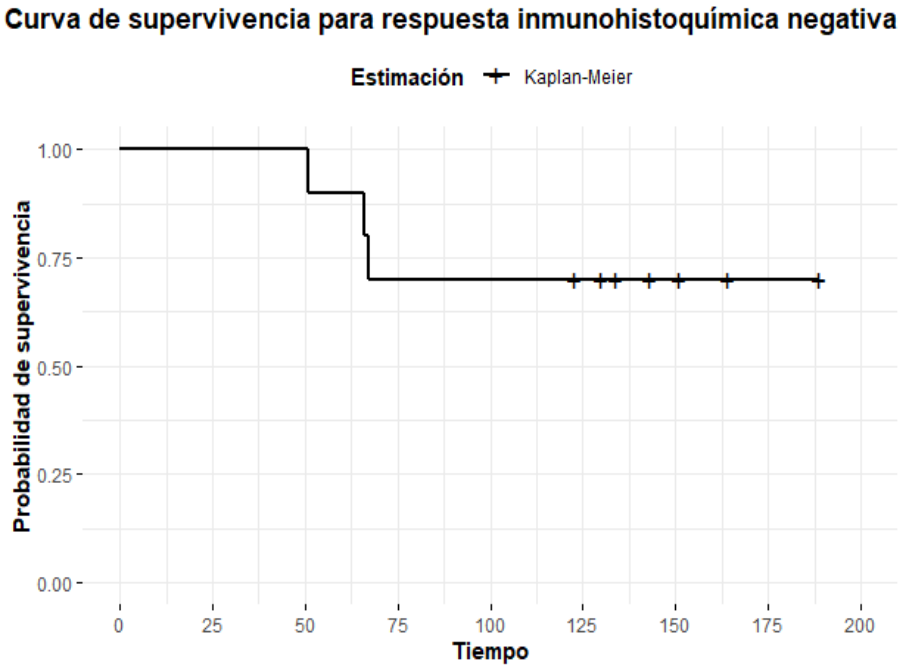


Gráfico 14. *Curva de supervivencia para una respuesta inmunohistoquímica negativa.*

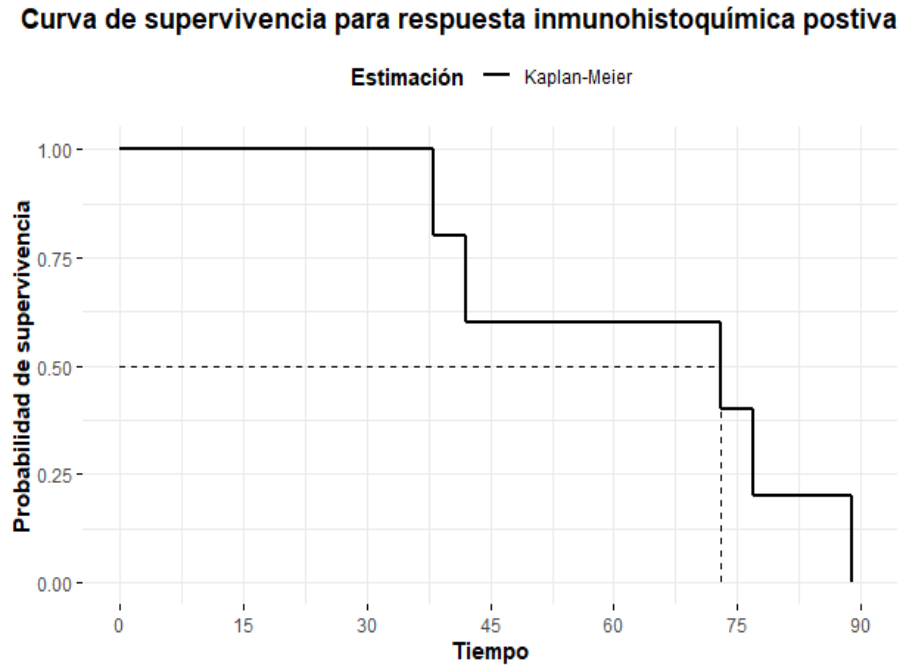


Gráfico 15. *Curva de supervivencia para una respuesta inmunohistoquímica positiva.*

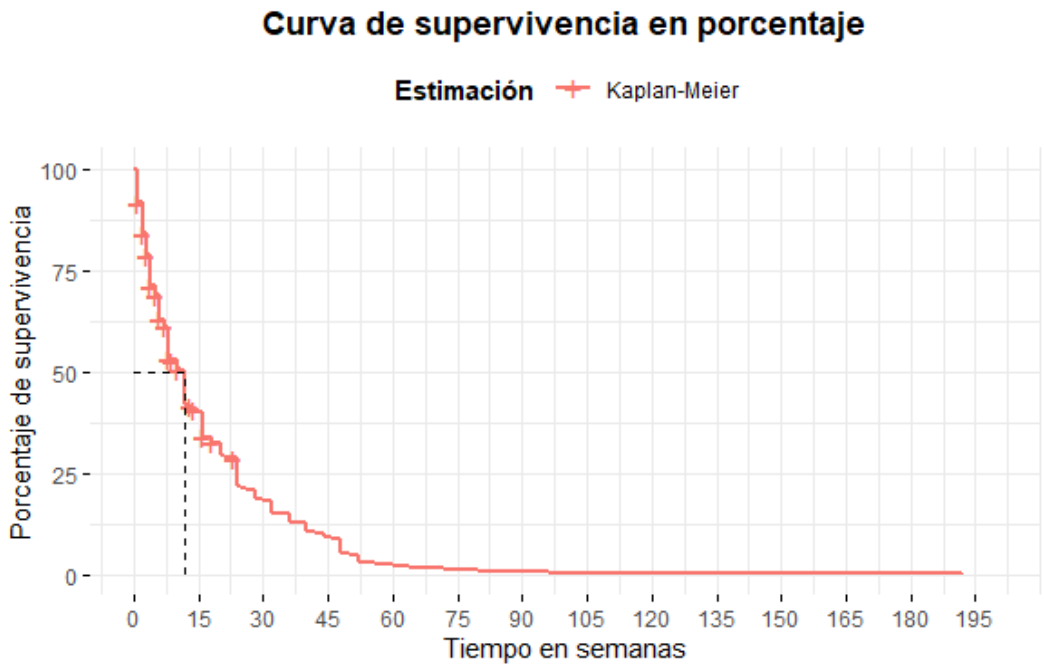


Gráfico 16. *Curva de supervivencia en porcentaje ("bfeed").*

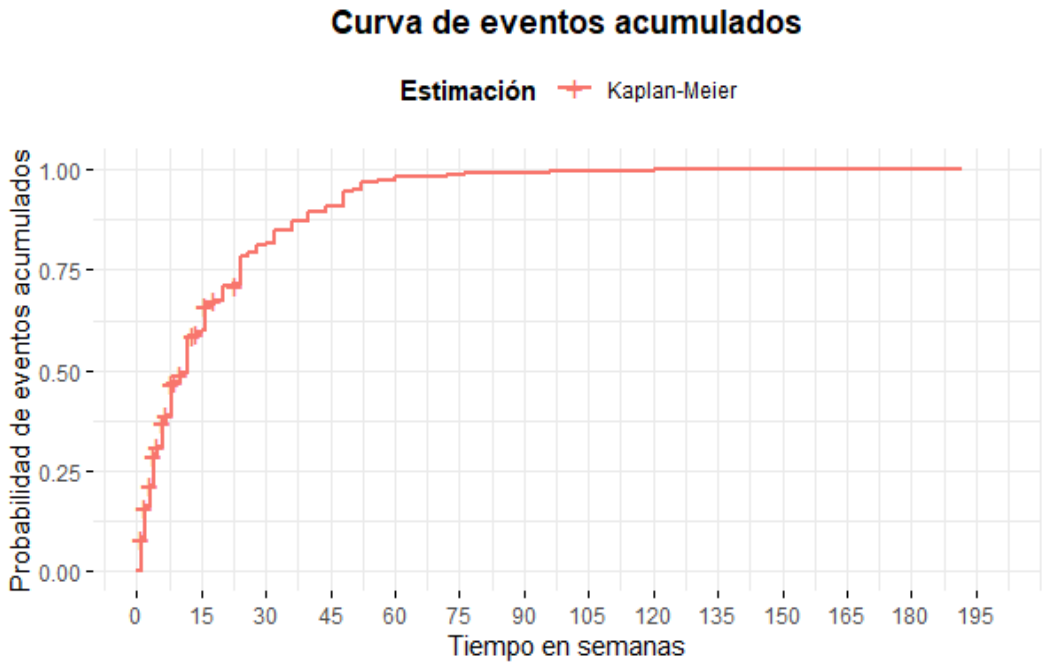


Gráfico 17. *Curva de eventos acumulados de la base de datos "bfeed".*