Universidad Autónoma Metropolitana
Unidad Iztapalapa

División de Ciencias Básicas e Ingeniería
Posgrado en Ciencias (Matemáticas Aplicadas e Industriales)

"Modelado de datos de supervivencia con una fracción no susceptible"

T E S I S

Que para obtener el grado de
Maestra en Ciencias
(Matemáticas Aplicadas e Industriales)

Presenta
Claudia Chaparro Velázquez
Matrícula: 2233803393
Correo electrónico: claudia.chaparro1998@gmail.com

Director de Tesis
Dr. Gabriel Escarela Pérez

Jurado

Presidente:	Dr. Gabriel Núñez Antonio
Secretario:	Dr. Gabriel Escarela Pérez
Vocal:	Dra. Elizabeth González Estrada

Iztapalapa, Ciudad de México 15 de Diciembre de 2025.

"Los estadísticos, como los artistas, tienen el mal hábito de enamorarse de sus modelos."

— George Box

"No fracasamos. Aprendimos con certeza que eso no funcionaba, y que teníamos que probar otra cosa."

— Thomas A. Edison

Agradecimientos

Agradezco profundamente al Dr. Gabriel Escarela Pérez, mi asesor de tesis, por haberme recibido en este proyecto, por compartir conmigo su experiencia, su apoyo constante, sus valiosos consejos y nuestras enriquecedoras pláticas, que me impulsaron a querer aprender más y a mirar siempre más allá.

Al Dr. Joaquín Delgado, coordinador del posgrado, gracias por su guía a lo largo de este viaje de maestría, por sus llamados de atención cuando mi camino no era claro, y por ayudarme a mantener el rumbo.

A la Universidad Autónoma Metropolitana, mi institución, gracias por brindarme la oportunidad de ser parte de este proyecto y de esta comunidad académica que me ha formado.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), gracias por el apoyo económico que me permitió dedicarme de lleno a mis estudios y continuar este camino académico.

A mis padres y a mi hermana, por su amor incondicional, por estar siempre ahí con su apoyo y palabras de aliento, incluso en los momentos más difíciles.

A Gabriel, gracias por creer en mí incluso cuando yo dudaba, por escucharme, por alentarme y recordarme que sí podía.

A Daniel, gracias por acompañarme en esta aventura, por tu compañía firme, tu amistad sincera y tu apoyo constante.

A mis compañeros y amigos: David, César, Abner, Antonio, Sara, y a todos con quienes compartí esta etapa, gracias por permitirme vivir momentos únicos llenos de estrés, risas, clases y aprendizaje. Ustedes hicieron de este proceso una experiencia más ligera, más humana y más inolvidable.

Y a mí misma: gracias por no rendirte, por levantarte cada vez que las dudas y la frustración quisieron detenerte. Por tu entrega, tu constancia y tu fortaleza silenciosa. Por seguir, incluso cuando el camino parecía nublado. Que este logro te recuerde siempre que eres capaz, que puedes lograr lo que te propongas, y que dentro de ti está la fuerza para superar cualquier reto. Cree en ti, siempre.

A todos ustedes, y a mí, gracias de corazón.

Resumen

En los estudios de supervivencia aplicados a contextos clínicos, el análisis del tiempo hasta la ocurrencia de un evento, como la recaída de una enfermedad, es una herramienta clave para la toma de decisiones médicas. No obstante, en ciertas enfermedades como el cáncer de mama, se ha observado que una proporción de pacientes no experimenta recaídas, incluso después de largos periodos de seguimiento. Este fenómeno sugiere la existencia de una fracción curada o no susceptible, lo que limita la aplicación directa de los modelos clásicos de supervivencia, ya que dichos modelos asumen que toda la población eventualmente sufrirá el evento si el seguimiento es lo suficientemente prolongado.

Cuando existe una fracción de individuos no susceptibles, este supuesto es incorrecto y provoca sesgos importantes pues se sobreestima el riesgo a largo plazo y se subestiman las probabilidades de supervivencia, distorsionando así el efecto de las covariables. Para abordar adecuadamente esta característica, se requieren modelos estadísticos que incorporen explícitamente la posibilidad de curación.

Este proyecto tiene como objetivo el desarrollo, ajuste e interpretación de un modelo de supervivencia con fracción curada, basado en un enfoque de mezcla que incorpora explícitamente los componentes de incidencia y latencia. La metodología considera, por un lado, un modelo logístico para estimar la probabilidad de que una paciente sea susceptible al evento (incidencia) y, por otro, un modelo de supervivencia con distribución Weibull para describir el tiempo hasta la recaída entre las pacientes susceptibles (latencia). El análisis se centra en un marcador génico continuo, cuyo nivel de expresión puede influir tanto en la probabilidad de ser susceptible como en la dinámica temporal del evento, permitiendo evaluar su efecto de manera integral sobre ambos componentes del proceso.

Para ilustrar la aplicación de esta metodología, se utiliza la base de datos pública *GSE2034*, disponible en el repositorio Gene Expression Omnibus (GEO). Esta base contiene información clínica y datos de expresión génica de pacientes con cáncer de mama en estadio temprano, incluyendo variables como el tiempo hasta la recaída, el estatus del evento, y perfiles de expresión de diversos genes. A partir de esta base, se seleccionó un subconjunto de datos con la información necesaria para construir el modelo, centrándose en un marcador génico de interés previamente identificado como potencialmente relevante en la progresión de la enfermedad.

El modelo se ajustó mediante un procedimiento de máxima verosimilitud, implementado en el software `R`, y se evaluó mediante residuos cuantílicos aleatorizados (RQR) y pruebas de normalidad (como la de Anderson-Darling). Los resultados indican que tanto en la probabilidad de ser susceptible (incidencia) como en la dinámica del tiempo hasta la recaída entre las pacientes susceptibles (latencia) la influencia del marcador es relevante; Específicamente, valores más altos del marcador se asocian con una menor probabilidad de ser susceptible al evento y para aquellos individuos que sí son susceptibles al evento, niveles altos del marcador aumentan significativamente el tempo de supervivencia antes de que ocurra el evento. La evaluación de residuos respalda la adecuación del modelo.

En conjunto, el modelo permite estimar simultáneamente la probabilidad de ser susceptible y la distribución del tiempo hasta el evento entre los individuos susceptibles, proporcionando una descripción más realista de la dinámica de supervivencia en poblaciones con fracción curada.

Índice general

Índice de figuras

Índice de tablas

Índice de algoritmos

Capítulo 1

Introducción

El análisis de supervivencia es un conjunto de métodos estadísticos enfocados en el estudio del tiempo hasta la ocurrencia de un evento de interés, como la muerte o la recurrencia de una enfermedad. En investigaciones médicas, frecuentemente se encuentran datos que representan el tiempo transcurrido hasta la ocurrencia de un evento particular, denominados datos de supervivencia. Estos datos presentan características especiales, como la censura, que ocurre cuando la información sobre el tiempo hasta el evento de interés está incompleta. Debido a estas particularidades, los datos de supervivencia no deben ser analizados mediante procedimientos estadísticos estándar [5, 18].

Uno de los métodos más comunes en el análisis de supervivencia es el estimador de Kaplan-Meier, que permite estimar la función de supervivencia de manera no paramétrica [10]. Además, el modelo de riesgos proporcionales de Cox es una extensión de este enfoque que permite analizar el efecto de varias covariables sobre el tiempo hasta el evento [6]. Sin embargo, ambos métodos tienen limitaciones. El estimador de Kaplan-Meier no permite analizar la influencia de covariables, mientras que el modelo de Cox asume que las razones de riesgos son proporcionales y constantes a lo largo del tiempo, lo cual no siempre se cumple en los datos reales [5].

Los modelos tradicionales de supervivencia también asumen que, dado suficiente tiempo, todos los individuos eventualmente experimentarán el evento de interés. No obstante, esta suposición puede ser inapropiada en situaciones donde existe evidencia de que una fracción de la población no está en riesgo. En el contexto del cáncer de mama, algunos pacientes pueden considerarse efectivamente curados después del tratamiento inicial, lo que sugiere la necesidad de modelos que incorporen explícitamente esta característica. Esta heterogeneidad en el riesgo debe modelarse adecuadamente para evitar estimaciones sesgadas y conclusiones imprecisas [11, 15].

Para superar estas limitaciones y proporcionar un análisis más preciso y realista en estudios de supervivencia, este proyecto propone el desarrollo y uso de un modelo de supervivencia con una fracción no susceptible. Este tipo de modelos permite tener en cuenta que una proporción de los individuos podría no experimentar nunca el evento de interés, incluso si se les observa durante un periodo prolongado de tiempo [11, 15].

Esta propuesta será ilustrada mediante datos de pacientes con cáncer de mama, ya que, tras el tratamiento inicial, algunos de ellos pueden considerarse efectivamente curados. En estos casos, es fundamental distinguir entre quienes aún están en riesgo de recaída y quienes no, con el fin de modelar adecuadamente su comportamiento a lo largo del tiempo. Ignorar esta distinción puede conducir a una subestimación de la supervivencia real y a interpretaciones erróneas [15, 14].

Además, se tendrá en cuenta el concepto de *heterogeneidad*, que hace referencia a la variabilidad no observada entre individuos. Esta puede deberse a factores clínicos, genéticos o ambientales que no han sido medidos o incluidos explícitamente en el modelo. Su presencia puede generar una mezcla de subpoblaciones con distintos riesgos de experimentar el evento, lo que influye directamente en la forma de la función de supervivencia [9].

1.1. Justificación

El uso de modelos de curación permite mejorar la precisión en la estimación de riesgos y tiempos esperados hasta la recaída, ofreciendo herramientas valiosas para la estratificación de pacientes y la personalización de tratamientos. Además, el análisis conjunto de dos componentes clave del proceso clínico—la incidencia (probabilidad de recaída) y la latencia (tiempo hasta la recaída entre los susceptibles)—proporciona una visión más completa y realista de la evolución del paciente. Estos modelos pueden ser especialmente útiles en oncología, donde la presencia de pacientes curados es clínicamente relevante.

1.2. Objetivos

Objetivo general:

Construir, ajustar e interpretar un modelo de supervivencia con fracción curada y aplicarlo a datos de pacientes con cáncer de mama en estadio temprano.

Objetivos específicos:

- Describir los fundamentos teóricos de los modelos de supervivencia y de curación.
- Desarrollar y aplicar un modelo de mezcla compuesto por un modelo logístico para la incidencia y un modelo de supervivencia Weibull para la latencia.
- Evaluar el ajuste del modelo mediante residuos cuantílicos aleatorizados y pruebas de normalidad.
- Analizar el efecto de un marcador molecular continuo sobre ambos componentes del modelo.

1.3. Hipótesis

Se plantea que el marcador molecular de interés tiene un efecto significativo tanto en la probabilidad de recaída como en el tiempo hasta la recaída, y que dicho efecto puede modelarse mediante una estructura de mezcla adecuada.

1.4. Estructura del documento

Este documento se organiza en varios capítulos. En los capítulos 2 a 6 se presentan los fundamentos teóricos de los modelos de supervivencia y de cura, incluyendo definiciones, formulaciones y supuestos principales. El capítulo 7 aborda los métodos para evaluar el ajuste del modelo mediante herramientas diagnósticas. El capítulo 8 describe el desarrollo y construcción del modelo, los procedimientos de estimación, así como la interpretación de resultados y la validación del modelo ajustado. Finalmente, en el capítulo 9 se discuten las implicaciones clínicas de los hallazgos y posibles líneas de investigación futura relacionadas con el uso de biomarcadores como herramienta para la estratificación de pacientes y la medicina personalizada.

Capítulo 2

Análisis de supervivencia

2.1. Conceptos básicos

El análisis de supervivencia, también conocido como análisis de datos en función del tiempo, es una rama de la estadística que se enfoca en estudiar el tiempo transcurrido hasta la ocurrencia de un evento de interés, como la recaída de una enfermedad o el fallecimiento de un paciente. Para ello, se define un punto temporal de origen (normalmente el momento en que un individuo ingresa al estudio) y un punto final en el que ocurre el evento de interés. Los datos obtenidos se denominan datos de supervivencia [5, 6, 8, 18].

Estos datos tienen características particulares que impiden analizarlos con métodos estadísticos clásicos [6]. Primero, presentan distribuciones asimétricas con sesgo positivo, lo que puede observarse en histogramas con una cola alargada hacia la derecha. Por ello, no es adecuado suponer que siguen una distribución normal [5, 6].

Otra razón fundamental es la presencia de censura en los tiempos de supervivencia, que ocurre cuando el evento de interés no ha sido observado durante el período de seguimiento. Por ejemplo, si un paciente se encuentra vivo en el último control sin haber experimentado el evento (como la muerte o recaída), su tiempo se considera *censurado a la derecha*, pues sólo se sabe que el evento no ocurrió hasta ese momento [5, 8, 11].

Existen otros tipos de censura:

- *Censura a la izquierda:* cuando el evento ocurrió antes de que comenzara la observación.
- *Censura por intervalos:* cuando sólo se sabe que el evento ocurrió en un intervalo de tiempo, pero no el instante exacto [5].

En este contexto, algunos términos importantes son:

- *Tiempo de estudio:* período en que el individuo está incluido en el seguimiento.
- *Tiempo del paciente:* duración desde la inclusión del paciente hasta el final del seguimiento o evento.
- *Tiempo de supervivencia:* intervalo entre el origen temporal y la ocurrencia del evento [5, 6, 11].

Por lo general, se asume que todos los individuos experimentarán el evento en algún momento, y se denota por T a la variable aleatoria continua que representa este tiempo [5]. Así mismo como norma general, nos centraremos en el estudio de un único evento [11].

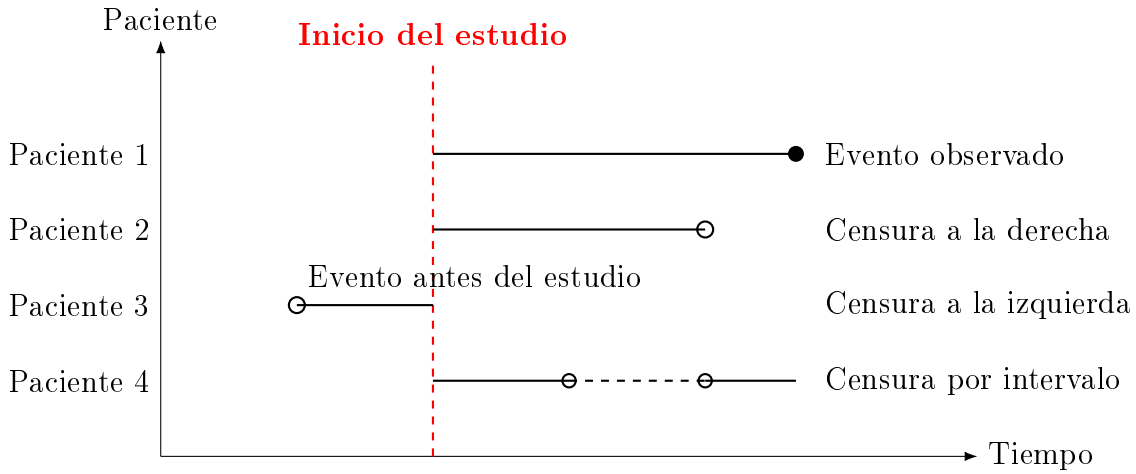


Figura 2.1: Representación de diferentes tipos de censura en datos de supervivencia.

2.2. Función de supervivencia

La *función de supervivencia* $S(t)$ es la probabilidad de que un individuo sobreviva más allá de un tiempo t, es decir, $S(t) = P(T > t)$. Dado que $F(t)$ es la función de distribución acumulada de T y $f(t)$ su función de densidad, entonces:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)\, du. \tag{2.1}$$

Donde:

- T representa el tiempo de supervivencia como una variable aleatoria.
- t es un instante fijo donde calculamos la probabilidad de que T sea mayor que t.

2.2.1. Propiedades de la Función de Supervivencia

La función de supervivencia es una función no creciente, lo que significa que a medida que t aumenta, $S(t)$ no puede aumentar. A continuación, analizamos dos casos importantes para el modelo de análisis de supervivencia básico:

- *Caso 1: Evaluación en $t = 0$*

  Sabemos que la función de distribución acumulada es:

  $$F(t) = P(T \leq t).$$

  Evaluando en $t = 0$:

  $$F(0) = P(T \leq 0) = 0.$$

  Esto implica que ningún individuo experimenta el evento en el instante inicial. Por lo tanto, la función de supervivencia en $t = 0$ es:

  $$S(0) = 1 - F(0) = 1 - 0 = 1.$$

  Esto nos indica que en el instante inicial t_0, la probabilidad de que un individuo haya experimentado el evento de interés (por ejemplo, fallecimiento) es cero, lo que implica que la probabilidad de continuar vivo en ese instante es uno [6].

- *Caso 2: Evaluación en $t \to \infty$*

  La función de distribución acumulada cumple la siguiente propiedad:

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

  Esto significa que, a medida que t aumenta indefinidamente, la probabilidad de que el evento haya ocurrido en algún momento antes de t tiende a 1.

  Ahora, evaluamos el límite de la función de supervivencia cuando $t \to \infty$:

  $$\lim_{t \to \infty} S(t) = \lim_{t \to \infty} (1 - F(t)) = 1 - 1 = 0.$$

  Esto indica que, si esperamos un tiempo suficientemente largo, la probabilidad de que el individuo siga vivo (o que el evento aún no haya ocurrido) tiende a cero. Es decir, eventualmente, todos los individuos experimentarán el evento de interés [5, 11].

2.3. Fuerza de mortalidad o función de riesgo

La *función de Riesgo* $h(t)$, representa la probabilidad instantánea de que un individuo experimente el evento de interés en un tiempo t, dado que ha "sobrevivido" hasta ese momento. En otras palabras, esta función indica la tasa de ocurrencia de un evento en un instante específico, condicionado a que el evento no haya ocurrido antes [6, 8].

2.3.1. Definición matemática

La función de riesgo $h(t)$ se define como:

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

Aplicando la definición de probabilidad condicional, se tiene:

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

Recordando que $P(t \leq T < t + \Delta t) = F(t + \Delta t) - F(t)$, donde $F(t)$ es la función de distribución acumulada, y que $P(T \geq t) = S(t) = 1 - F(t)$, con $S(t)$ la función de supervivencia, podemos reescribir:

$$h(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \cdot \frac{1}{S(t)}$$

El primer término es, por definición, la derivada de la función de distribución acumulada:

$$\lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} = f(t)$$

donde $f(t)$ es la función de densidad de probabilidad.

Por lo tanto, obtenemos la expresión final de la función de riesgo:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.2}$$

2.3.2. Función de Riesgo Acumulado

La función de riesgo acumulado $H(t)$ es fundamental para modelos de supervivencia. Esta función mide la cantidad total de riesgo acumulado hasta el momento t y se define como:

$$H(t) = \int_0^t h(x)dx = \int_0^t \frac{f(x)}{S(x)}dx = \int_0^t \frac{F'(x)}{1-F(x)}dx \tag{2.3}$$

de donde se obtiene la expresión:

$$H(t) = -\log(1-F(t)) = -\log(S(t)) \tag{2.4}$$

Que a su vez nos permite expresar la función de supervivencia en términos de la función de riesgo acumulado:

$$S(t) = e^{-H(t)} \tag{2.5}$$

Estas funciones permiten comprender cómo varía el riesgo a lo largo del tiempo y son fundamentales para la modelación estadística en supervivencia [5, 6].

2.4. Función de Verosimilitud

Para estimar los parámetros de un modelo de supervivencia se construye la función de verosimilitud basada en los tiempos observados t_i donde δ_i es un indicador que vale 0 si el dato está censurado y 1 si el evento fue observado para el sujeto i, cuya función de verosimilitud es:

$$L(\theta) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{2.6}$$

Su logaritmo es:

$$\log L(\theta) = \sum_{i=1}^{n} [\delta_i \log f(t_i) + (1-\delta_i) \log S(t_i)] \tag{2.7}$$

Puede expresarse en función del riesgo $h(t)$ y la supervivencia $S(t)$:

$$L = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i) \tag{2.8}$$

y tomando el logaritmo obtenemos la función de log-verosimilitud:

$$\ell = \sum_{i=1}^{n} [\delta_i \log h(t_i) + \log S(t_i)] \tag{2.9}$$

Esta función de verosimilitud es la base para la estimación por máxima verosimilitud en los modelos paramétricos de supervivencia que se abordarán más adelante [5, 8, 11].

Capítulo 3

Estimadores no paramétricos

En el análisis de supervivencia, los estimadores no paramétricos permiten estudiar la distribución del tiempo hasta un evento sin asumir una forma funcional específica. Este enfoque es especialmente útil cuando se desconoce la distribución subyacente o cuando se desea una descripción preliminar de los datos.

En este capítulo se presentan tres herramientas fundamentales: el estimador de Kaplan-Meier, el estimador de Nelson-Aalen y la prueba log-rank.

3.1. Estimador de Kaplan-Meier

El estimador de Kaplan-Meier es una herramienta esencial para estimar la función de supervivencia $S(t)$ en presencia de censura [10, 18]. Su formulación matemática, basada en probabilidades condicionales, lo hace robusto y flexible.

Este estimador se construye a partir de los tiempos observados de eventos y los tiempos censurados. Su objetivo es estimar la probabilidad de supervivencia en cada tiempo t utilizando la información disponible hasta ese momento.

3.1.1. Definición matemática del Estimador KM

El estimador de Kaplan-Meier puede formularse matemáticamente de maneras equivalentes, las cuales se muestran a continuación:

- *Basado en los tiempos de recaída observados*

  Supongamos que se tienen N eventos de recaída observados, ordenados de menor a mayor: $t'_1 < t'_2 < \ldots < t'_N$. El estimador de Kaplan-Meier se puede expresar como:

  $$\hat{S}(t) = \prod_{t'_r \leq t} \left(\frac{N - r}{N - r + 1} \right)$$

  donde r es el índice de los tiempos de recaída observados antes o iguales a t.

- *Como producto de probabilidades condicionales de supervivencia*

  Supongamos que se tienen n individuos en riesgo al inicio del estudio. Los tiempos de eventos se ordenan de menor a mayor: $t_1 < t_2 < \ldots < t_k$.

  Para cada tiempo t_j, se definen:

  - n_j: Número de individuos en riesgo justo antes del tiempo t_j.
  - d_j: Número de eventos (recaídas) observados en el tiempo t_j.

  La probabilidad condicional de supervivencia en el intervalo $(t_{j-1}, t_j]$ se estima como:

$$p_j = \frac{n_j - d_j}{n_j}$$

  El estimador de Kaplan-Meier de la función de supervivencia $S(t)$ en el tiempo t es el producto de estas probabilidades condicionales para todos los tiempos t_j menores o iguales a t:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j} \right) \tag{3.1}$$

3.1.2. Propiedades del Estimador de Kaplan-Meier

- El estimador de Kaplan-Meier es una función escalonada que cambia su valor solo en los tiempos en que se observan los eventos. En estos puntos, la función tiene una discontinuidad [6, 10].
- Entre dos tiempos de eventos consecutivos, el estimador de la función de supervivencia permanece constante.
- El estimador no asume ninguna forma funcional específica para la distribución del tiempo hasta el evento, lo que lo hace flexible y robusto en situaciones donde la distribución subyacente es desconocida. En otras palabras, resulta un método no paramétrico.

3.1.3. Suposiciones Clave

El uso del estimador de Kaplan-Meier implica suponer que:

- La censura es independiente del tiempo de supervivencia.
- Los individuos censurados tienen la misma probabilidad de supervivencia que los no censurados hasta ese momento.

3.1.4. Aplicaciones

El estimador de Kaplan-Meier proporciona una base para inferencias estadísticas sobre el tiempo hasta el evento, como la mediana de supervivencia o la probabilidad de supervivencia a un tiempo específico. Además, es ampliamente utilizado en investigación médica:

- Para estimar la supervivencia de pacientes después de un tratamiento o diagnóstico.
- Para analizar el tiempo hasta el fallo.
- Comparación de Grupos: Para evaluar diferencias en la supervivencia entre grupos, como pacientes que reciben diferentes tratamientos.

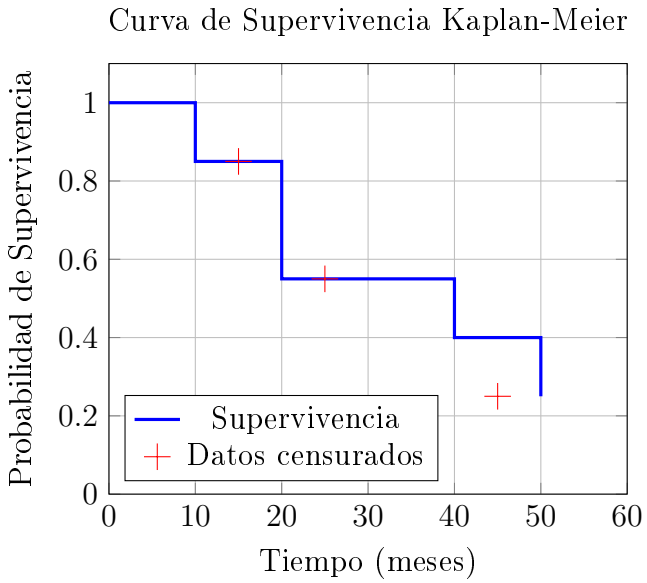


Figura 3.1: Curva de supervivencia con datos censurados (marcas +). Cada escalón representa un evento de recaída.

3.2. Estimador de Nelson-Aalen

El estimador de Nelson-Aalen fue propuesto de manera independiente por Nelson (1972) y Aalen (1978) como una alternativa no paramétrica utilizada para estimar la función de riesgo acumulado $H(t)$ a lo largo del tiempo. Este estimador es particularmente útil en contextos donde se desea analizar la tasa acumulada de ocurrencia de eventos sin asumir una forma funcional específica para la distribución del tiempo hasta el evento [18, 5].

3.2.1. Definición matemática:

La función de riesgo acumulado $H(t)$ se define como la integral de la función de riesgo instantáneo $h(t)$:

$$H(t) = \int_0^t h(u)\,du$$

Cuando trabajamos con datos observados (discretos), esta integral puede aproximarse mediante una suma de Riemann.

Supongamos que los eventos ocurren en los tiempos ordenados $t_1 < t_2 < \ldots < t_m$. Para cada tiempo t_i, definimos:

- d_i: número de eventos observados en el tiempo t_i.
- n_i: número de individuos en riesgo justo antes del tiempo t_i, es decir, aquellos que aún no han experimentado el evento ni han sido censurados antes de t_i.

A partir de estos valores, podemos estimar empíricamente el riesgo instantáneo en el tiempo t_i como:

$$\hat{h}(t_i) = \frac{d_i}{n_i}$$

Esta expresión representa la proporción de individuos en riesgo que experimentan el evento en el instante t_i.

Como la función de riesgo acumulado $H(t)$ es la integral de $h(t)$, en un contexto de datos discretos la aproximamos por la suma de los riesgos instantáneos observados hasta el tiempo t:

$$\hat{H}(t) = \sum_{t_i \leq t} \hat{h}(t_i) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{3.2}$$

Esta es la fórmula del *estimador de Nelson-Aalen*, que permite estimar de manera no paramétrica el riesgo acumulado hasta un tiempo t. Se trata de una herramienta fundamental en el análisis de supervivencia cuando no se desea suponer una forma funcional específica para la función de riesgo [6].

Ejemplo ilustrativo del estimador de Nelson-Aalen

Tabla 3.1: Tiempos de ocurrencia de eventos de falla en la muestra

Tiempo t_i	d_i (eventos)	n_i (en riesgo)	$\hat{h}(t_i) = \frac{d_i}{n_i}$
1	1	10	0.100
2	1	9	0.111
3	2	8	0.250
5	1	6	0.167

La estimación del riesgo acumulado en el tiempo $t = 5$ se obtiene sumando las proporciones:

$$\hat{H}(5) = \frac{1}{10} + \frac{1}{9} + \frac{2}{8} + \frac{1}{6} \approx 0.100 + 0.111 + 0.250 + 0.167 = 0.628$$

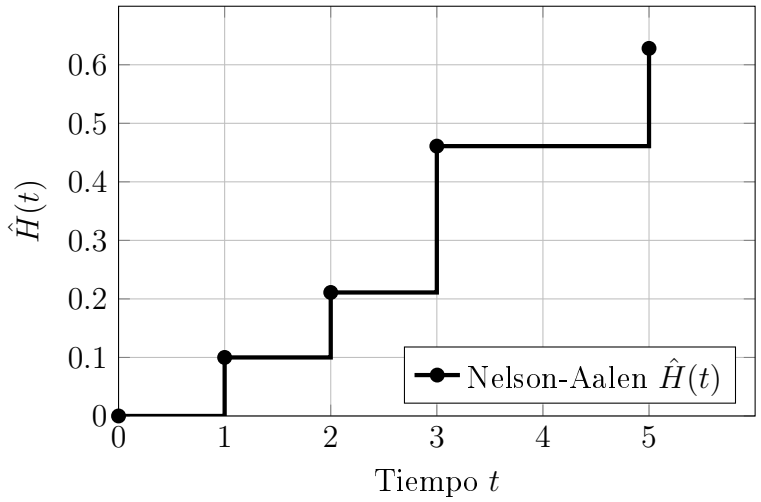


Figura 3.2: Estimación de la función de riesgo acumulado con el estimador de Nelson-Aalen.

La curva muestra la acumulación del riesgo a lo largo del tiempo. En cada tiempo t_i donde ocurre al menos un evento, se calcula la fracción $\frac{d_i}{n_i}$, que representa el riesgo instantáneo en ese momento. Luego, se acumulan estas fracciones de manera secuencial:

$$\hat{H}(1) = \frac{1}{10} = 0.100$$
$$\hat{H}(2) = \hat{H}(1) + \frac{1}{9} = 0.211$$
$$\hat{H}(3) = \hat{H}(2) + \frac{2}{8} = 0.461$$
$$\hat{H}(5) = \hat{H}(3) + \frac{1}{6} = 0.628$$

La curva conecta estos puntos con líneas rectas para representar visualmente cómo se acumula el riesgo. En realidad, el estimador tiene una forma de *escalones*, ya que el riesgo acumulado solo cambia en los tiempos donde ocurren eventos.

3.2.2. Consideraciones sobre el uso del estimador de Nelson-Aalen

En muchos estudios de supervivencia, los tiempos de ocurrencia del evento (por ejemplo, recaída, muerte, falla de un sistema) son en teoría variables continuas. Sin embargo, en la práctica, estos tiempos se registran en unidades discretas (como días, semanas o meses), lo que permite aplicar directamente métodos no paramétricos como el estimador de Nelson-Aalen.

Incluso si los tiempos de falla siguen una distribución continua desconocida, el estimador de Nelson-Aalen sigue siendo válido, ya que:

- No requiere asumir una distribución paramétrica para los tiempos de falla.
- Utiliza únicamente los tiempos en los que efectivamente se observan eventos.
- Se construye empíricamente a partir de proporciones observadas $\frac{d_i}{n_i}$, por lo que es robusto frente a supuestos de forma funcional.

Por lo tanto, el estimador de Nelson-Aalen es una herramienta fundamental del enfoque no paramétrico en el análisis de supervivencia, y resulta especialmente útil cuando:

- No se desea imponer una distribución específica.
- Se busca una estimación robusta del riesgo acumulado basada únicamente en los datos observados.

En contraste, si se conoce o se asume una forma funcional específica para la distribución del tiempo hasta el evento (por ejemplo, exponencial, Weibull o log-normal), entonces pueden usarse modelos paramétricos que proporcionan estimaciones más eficientes [5, 8, 18].

3.3. Prueba Log-Rank

Mantel y Haenszel (1959) propusieron la prueba de Log-Rank como herramienta estadística utilizada para comparar las curvas de supervivencia entre dos o más grupos [5, 18]. Es especialmente útil en el análisis de supervivencia, donde se evalúa el tiempo hasta que ocurre un evento de interés, como puede ser la recaída de una enfermedad, el fallo de un dispositivo o el fallecimiento de un paciente. Esta prueba es no paramétrica y se basa en la hipótesis nula H_0 de que no existen diferencias significativas entre las curvas de supervivencia de los grupos comparados, es decir, $S_1(t) = S_2(t)$ para todo t.

Supongamos que deseamos comparar la supervivencia entre dos grupos (por ejemplo, pacientes tratados con dos fármacos distintos). En cada tiempo en que se observa un evento (digamos, una recaída), la prueba analiza si el número de eventos en cada grupo coincide con el número esperado de eventos, bajo la hipótesis de que ambos grupos tienen la misma probabilidad de supervivencia. Cuando los eventos observados difieren sistemáticamente de los esperados, esto sugiere que existe una diferencia real entre los grupos.

3.3.1. Formulación matemática

Sean dos grupos de individuos, el grupo 1 y el grupo 2. En cada tiempo t_j donde ocurre al menos un evento, se definen las siguientes cantidades:

- n_{1j}: Número de individuos en riesgo en el grupo 1 justo antes de t_j.
- n_{2j}: Número de individuos en riesgo en el grupo 2 justo antes de t_j.
- d_{1j}: Número de eventos observados en el grupo 1 en t_j.
- d_{2j}: Número de eventos observados en el grupo 2 en t_j.
- $n_j = n_{1j} + n_{2j}$: Total de individuos en riesgo justo antes de t_j.
- $d_j = d_{1j} + d_{2j}$: Total de eventos observados en t_j.

Queremos comparar los valores observados d_{1j} y d_{2j} con sus valores esperados. Bajo la hipótesis nula H_0, los eventos deberían distribuirse proporcionalmente al número de individuos en riesgo en cada grupo. Así, el número esperado de eventos en el grupo 1 en el tiempo t_j es:

$$E_{1j} = \frac{n_{1j} \cdot d_j}{n_j}$$

y en el grupo 2:

$$E_{2j} = \frac{n_{2j} \cdot d_j}{n_j}$$

De este modo, se cumple que:

$$E_{1j} + E_{2j} = d_j$$

Para evaluar la diferencia en supervivencia, se analizan las diferencias entre los valores observados y esperados:

$$O_{1j} - E_{1j} = d_{1j} - \frac{n_{1j} \cdot d_j}{n_j}$$

$$O_{2j} - E_{2j} = d_{2j} - \frac{n_{2j} \cdot d_j}{n_j}$$

Si estas diferencias son grandes en valor absoluto, puede indicar una diferencia significativa en las curvas de supervivencia.

El número de eventos d_{1j} puede modelarse mediante una distribución hipergeométrica, ya que se seleccionan eventos sin reemplazo de una población finita de n_j individuos en riesgo. La varianza de una variable hipergeométrica es:

$$V_{1j} = E_{1j} \cdot \left(1 - \frac{n_{1j}}{n_j}\right) \cdot \frac{n_j - d_j}{n_j - 1}$$

Al expresarla completamente en función de las variables originales, se obtiene:

$$V_{1j} = \frac{n_{1j} \cdot n_{2j} \cdot d_j \cdot (n_j - d_j)}{n_j^2 \cdot (n_j - 1)}$$

La varianza total acumulada se obtiene sumando sobre todos los tiempos en que ocurre un evento:

$$V = \sum_j V_{1j}$$

3.3.2. Estadístico Log-Rank

El estadístico Log-Rank compara las diferencias acumuladas entre eventos observados y esperados, y se define como:

$$Z = \frac{\sum_j (O_{1j} - E_{1j})}{\sqrt{\sum_j V_{1j}}} \tag{3.3}$$

Este estadístico sigue aproximadamente una distribución normal estándar bajo H_0. Un valor de Z cercano a cero sugiere que los grupos presentan patrones similares de supervivencia. En cambio, un valor elevado (positivo o negativo) sugiere que las diferencias observadas no se deben al azar, y puede rechazarse H_0 en favor de la hipótesis alternativa $H_A : S_1(t) \neq S_2(t)$.

3.3.3. Prueba de significancia

El cuadrado del estadístico Z sigue una distribución chi-cuadrado con 1 grado de libertad:

$$\chi^2 = Z^2 \tag{3.4}$$

Si el valor-p asociado a χ^2 es menor que un nivel de significancia preestablecido (por ejemplo, $p < 0.05$), se rechaza la hipótesis nula, concluyendo que las curvas de supervivencia difieren significativamente entre los grupos comparados.

En resumen, la prueba Log-Rank es una de las herramientas más empleadas en estudios de supervivencia, gracias a su simplicidad, robustez frente a censura no informativa, y aplicabilidad a múltiples contextos clínicos y de ingeniería.

Ejemplo ilustrativo de la prueba log-rank

Supongamos que comparamos la supervivencia entre dos grupos de pacientes: *Grupo 1* y *Grupo 2*, observando eventos (recaídas) en distintos tiempos. Los datos son los siguientes:

Tabla 3.2: Datos de comparación entre dos grupos para la prueba Log-Rank

Tiempo t_j	n_{1j}	n_{2j}	d_{1j}	d_{2j}
1	10	10	1	2
2	9	8	2	1
3	7	7	1	2

Paso 1. Calcular los eventos esperados para el Grupo 1 en cada t_j:

$$E_{1j} = \frac{n_{1j}}{n_j} \cdot d_j, \quad \text{donde } n_j = n_{1j} + n_{2j}, \quad d_j = d_{1j} + d_{2j}$$

- En t_1: $E_{11} = \frac{10}{20} \cdot 3 = 1.5$
- En t_2: $E_{12} = \frac{9}{17} \cdot 3 \approx 1.588$
- En t_3: $E_{13} = \frac{7}{14} \cdot 3 = 1.5$

Paso 2. Sumar las diferencias observadas menos esperadas:

$$\sum_j (O_{1j} - E_{1j}) = (1 - 1.5) + (2 - 1.588) + (1 - 1.5) = -0.588$$

Paso 3. Calcular la varianza en cada tiempo (usando la fórmula hipergeométrica):

$$V_{1j} = \frac{n_{1j} \cdot n_{2j} \cdot d_j \cdot (n_j - d_j)}{n_j^2 \cdot (n_j - 1)}$$

- $V_{11} \approx 0.395$
- $V_{12} \approx 0.468$
- $V_{13} \approx 0.396$

Entonces, la varianza total es:

$$V = \sum_j V_{1j} \approx 1.259$$

Paso 4. Calcular el estadístico log-rank:

$$Z = \frac{-0.588}{\sqrt{1.259}} \approx -0.524$$

Este valor de Z indica que no hay evidencia suficiente para rechazar la hipótesis nula. Las diferencias entre los grupos podrían deberse al azar.

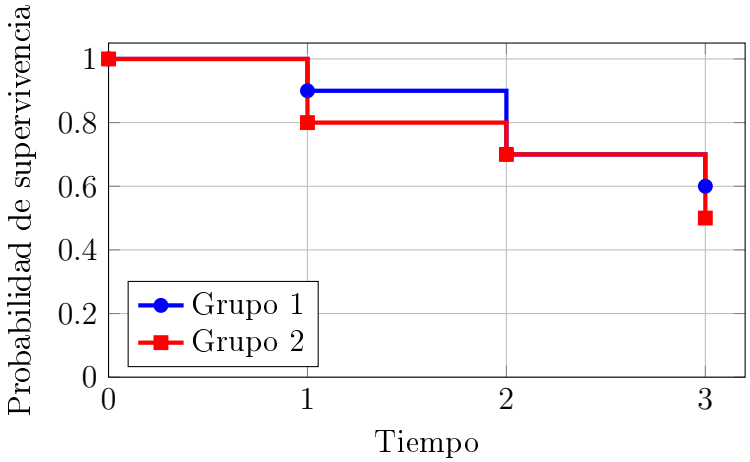


Figura 3.3: Curvas de supervivencia estimadas para los Grupos 1 y 2.

En la Figura 3.3 se observa que el Grupo 2 (línea roja) tiende a tener una menor probabilidad de supervivencia en cada punto temporal en comparación con el Grupo 1 (línea azul). Sin embargo, al aplicar la prueba log-rank, se obtiene un valor del estadístico $Z = -0.524$, el cual está dentro del rango de aceptación bajo la hipótesis nula H_0, que plantea que ambas curvas de supervivencia son iguales.

Por lo tanto, no se rechaza la hipótesis nula, y se concluye que no hay evidencia suficiente para afirmar que las curvas de supervivencia de los grupos 1 y 2 son diferentes.

Capítulo 4

Estimadores paramétricos

Cuando hablamos de análisis de supervivencia, el objetivo es estudiar el tiempo que pasa hasta que ocurre un evento, como la recaída de una enfermedad en un paciente. En este caso, los estimadores paramétricos nos ayudan a modelar cómo varía el riesgo de recaída a lo largo del tiempo, es decir, estos modelos suponen una forma funcional para la distribución de los tiempos hasta el evento y permiten hacer inferencias más eficientes cuando esa suposición es razonable [18, 5].

Dependiendo de la naturaleza del fenómeno, el riesgo puede mantenerse constante, aumentar o disminuir con el tiempo. Algunos modelos paramétricos comunes en el análisis de supervivencia son la distribución exponencial y la distribución de Weibull.

La distribución exponencial es uno de los modelos más simples y más utilizados en el análisis de supervivencia. Su principal característica es que asume una tasa de riesgo constante en el tiempo, lo que implica que la probabilidad instantánea de experimentar el evento no depende del tiempo transcurrido. Debido a esta propiedad, la distribución exponencial ha sido ampliamente empleada en estudios donde el riesgo permanece estable durante el seguimiento, como en el análisis de tiempos de falla de componentes electrónicos o dispositivos médicos [18]. También se ha utilizado en procesos biológicos que presentan una progresión aproximadamente uniforme en el tiempo, así como en estudios iniciales de recaída de enfermedades, donde resulta razonable asumir un riesgo estacionario antes de introducir modelos más complejos [5].

Sin embargo, el modelo exponencial limita su aplicabilidad en situaciones donde el riesgo varía con el tiempo, como lo es en el caso de enfermedades crónicas y en la progresión del cáncer. Para abordar esta limitación, se recurre con frecuencia a la distribución Weibull, que puede interpretarse como una generalización del modelo exponencial. La distribución Weibull permite modelar tasas de riesgo crecientes, decrecientes o constantes, dependiendo del valor de su parámetro de forma, lo que la convierte en una herramienta sumamente flexible para el análisis de supervivencia [5, 6].

Gracias a esta flexibilidad, la distribución Weibull es ampliamente utilizada en estudios clínicos, especialmente en cáncer, donde el riesgo de recaída puede incrementarse o disminuir conforme avanza el tiempo [11]. Además, su uso se extiende a áreas como la fiabilidad y la durabilidad, entre ellas la evaluación de tiempos de falla de prótesis, válvulas cardiovasculares y otros dispositivos implantables. Asimismo, la distribución Weibull se emplea en epidemiología para describir la duración de enfermedades crónicas y en modelos con heterogeneidad no observada, incluyendo modelos de mezcla y modelos de cura, debido a su capacidad para capturar variaciones del riesgo entre subpoblaciones [11].

En síntesis, mientras que la distribución exponencial resulta adecuada para situaciones donde el riesgo permanece constante, la distribución Weibull permite modelar de manera más realista procesos biológicos y clínicos en los que el riesgo evoluciona con el tiempo. Su uso en estudios de cáncer, fallas de dispositivos médicos y procesos epidemiológicos destaca su relevancia y versatilidad dentro del análisis de supervivencia.

4.1. Distribución Exponencial

La distribución exponencial supone que el riesgo de recaída es constante en el tiempo. Es útil cuando la probabilidad de recaer no cambia con el tiempo, es decir, cuando cada instante tiene la misma probabilidad de recaída. Se trata de un caso especial de la distribución de Weibull, donde el parámetro de forma $\alpha = 1$.

4.1.1. Función de densidad de probabilidad (PDF)

Está dada por:

$$f(t;\lambda) = \lambda e^{-\lambda t}, \quad t \geq 0$$

Donde:

- t: tiempo hasta la recaída
- λ: tasa de ocurrencia del evento (inverso de la media)

4.1.2. Función de distribución acumulada (CDF)

La probabilidad de que el evento ocurra antes de un cierto tiempo t se expresa como:

$$F(t, \lambda) = P(T \leq t) = \int_0^t \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_0^t = 1 - e^{-\lambda t}$$

La CDF de la distribución exponencial crece con t y se aproxima a 1 a medida que t aumenta, lo que significa que, con el paso del tiempo, la probabilidad de que el evento ocurra se va incrementando.

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

4.1.3. Función de supervivencia

La probabilidad de que el evento no haya ocurrido hasta el tiempo t se obtiene como:

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

Es complementaria a la función de distribución acumulada.

4.1.4. Función de riesgo

La función de riesgo (o tasa de fallos) es la tasa a la que ocurren los eventos en un intervalo de tiempo dado, condicionado a que no haya ocurrido el evento antes. Para la distribución exponencial, la función de riesgo es constante y está dada por:

$$h(t; \lambda) = \frac{f(t; \lambda)}{S(t; \lambda)} = \lambda$$

Esta es una característica distintiva de la distribución exponencial.

4.1.5. Parámetro de la distribución

La distribución exponencial tiene un solo parámetro, λ, que se interpreta como la tasa de ocurrencia del evento. Es el inverso de la media μ:

$$\lambda = \frac{1}{\mu}$$

- Si λ es grande, los eventos ocurren más rápidamente.
- Si λ es pequeño, los eventos ocurren menos frecuentemente.

4.2. Distribución de Weibull

La distribución de Weibull es una generalización de la distribución exponencial. Permite modelar riesgos que cambian con el tiempo, ya sea aumentando o disminuyendo. Es especialmente útil en situaciones donde se sospecha que la tasa de recaída no es constante[18, 5].

4.2.1. Función de densidad de probabilidad (PDF)

La función de densidad de Weibull, describe cómo se distribuyen los tiempos hasta que ocurre un evento. Es decir, nos indica la probabilidad de que el evento ocurra exactamente en un tiempo t, y está definida como:

$$f(t) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^{\alpha}}, \quad t > 0 \tag{4.1}$$

Donde:

- α*: Parámetro de forma*
  - Si $\alpha = 1$: riesgo constante (distribución exponencial)
  - Si $\alpha > 1$: el riesgo aumenta con el tiempo
  - Si $\alpha < 1$: el riesgo disminuye con el tiempo

- λ*: Parámetro de escala*
  - Controla la "escala temporal" de la distribución.
  - A mayor λ, los eventos tienden a ocurrir en tiempos más largos.

Interpretación de los componentes de la función:

- El término $\left(\frac{t}{\lambda}\right)^{\alpha-1}$ controla cómo cambia la probabilidad del evento con respecto al tiempo t.

- El término $e^{-\left(\frac{t}{\lambda}\right)^{\alpha}}$ representa la *probabilidad de supervivencia*, es decir, la probabilidad de que el evento *no haya ocurrido* hasta el tiempo t.
  - Si $\alpha > 1$, la supervivencia disminuye rápidamente con el tiempo.
  - Si $\alpha < 1$, la supervivencia disminuye más lentamente.

4.2.2. Función de distribución acumulada (CDF)

Se puede obtener directamente como la integral de la función de densidad de probabilidad:

$$F(t;\alpha,\lambda) = P(T \leq t) = \int_0^t f(s;\alpha,\lambda)\,ds = \int_0^t \frac{\alpha}{\lambda}\left(\frac{s}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{s}{\lambda}\right)^{\alpha}} ds$$

Otra forma más directa de obtener la CDF es reconociendo que:

$$\frac{d}{dt}\left(1 - e^{-\left(\frac{t}{\lambda}\right)^{\alpha}}\right) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^{\alpha}} = f(t;\alpha,\lambda)$$

Por lo tanto, la función de distribución acumulada de Weibull también se expresa como:

$$F(t;\alpha,\lambda) = 1 - e^{-\left(\frac{t}{\lambda}\right)^{\alpha}}, \quad t > 0 \tag{4.2}$$

4.2.3. Función de supervivencia

La función de supervivencia es la probabilidad de que el evento no haya ocurrido antes del tiempo t. Se define como:

$$S(t;\alpha,\lambda) = 1 - F(t;\alpha,\lambda) = e^{-\left(\frac{t}{\lambda}\right)^{\alpha}}, \quad t > 0 \tag{4.3}$$

4.2.4. Función de riesgo

La función de riesgo (hazard) representa la tasa instantánea de ocurrencia del evento en el tiempo t, dado que no ha ocurrido antes. Se calcula como el cociente entre la función de densidad $f(t)$ y la función de supervivencia $S(t)$:

$$h(t;\alpha,\lambda) = \frac{f(t;\alpha,\lambda)}{S(t;\alpha,\lambda)} = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}, \quad t > 0 \tag{4.4}$$

4.2.5. Aplicación en análisis de recaída

En el contexto del análisis de recaída por cáncer de mama, estos modelos permiten responder preguntas como: ¿la probabilidad de recaída cambia con el tiempo?, ¿existe un periodo más riesgoso?

Definimos:

- T: tiempo hasta la recaída.
- $f(t)$: densidad de la recaída en el tiempo t.
- $S(t)$: probabilidad de no haber recaído hasta el tiempo t.
- $h(t)$: riesgo instantáneo de recaída, dado que no ha ocurrido antes.

En resumen, la distribución de Weibull es fundamental por su flexibilidad para representar diferentes escenarios clínicos y temporales. Su aplicación permite entender mejor la evolución de enfermedades crónicas y orientar estrategias de tratamiento y seguimiento [18].

Capítulo 5

Modelo semiparamétrico

Un modelo semiparamétrico es un tipo de modelo estadístico que combina dos componentes:

- Un componente *paramétrico*, que depende de un número finito de parámetros (como en una regresión lineal, por ejemplo: $Y = \beta_0 + \beta_1 X$).
- Un componente *no paramétrico*, que no asume una forma funcional específica y se estima directamente a partir de los datos.

Esto significa que el modelo no especifica completamente la distribución de los datos, lo que le otorga una mayor flexibilidad en su aplicación. Al evitar la necesidad de asumir una forma funcional concreta, permite además interpretar los coeficientes β como en los modelos paramétricos [6, 5]. Gracias a estas características, los modelos semiparamétricos presentan un menor riesgo de especificación incorrecta, lo que los hace especialmente útiles en contextos donde no se conoce con certeza la estructura exacta del fenómeno que se desea modelar [5, 18, 8].

Dentro de este enfoque se encuentra el *modelo de riesgos proporcionales de Cox*, el cual es el modelo semiparamétrico más utilizado en análisis de supervivencia. Sin embargo, presenta una limitación fundamental: el supuesto de *riesgos proporcionales*. Este supuesto exige que el efecto de las covariables permanezca constante a lo largo del tiempo; cuando no se cumple, las estimaciones pueden ser sesgadas y su interpretación poco fiable. Por ello, es indispensable evaluar este supuesto en cada aplicación.

5.1. Modelo de Riesgos Proporcionales de Cox

El objetivo principal del modelo de riesgos proporcionales de Cox es describir cómo ciertas variables explicativas, llamadas *covariables*, afectan la tasa de ocurrencia del evento de interés, también conocida como función de riesgo o fuerza de mortalidad. En otras palabras, el modelo busca expresar la probabilidad instantánea de que ocurra un evento en el tiempo t, dado que el individuo ha sobrevivido hasta ese momento, en función del tiempo y de las covariables observadas [5].

5.1.1. Formulación matemática

Cox definió la función de riesgo o fuerza de mortalidad de un individuo con vector de covariables X y el correspondiente vector de coeficientes β, como:

$$h(t \mid X) = h_0(t)e^{\beta^T X} \tag{5.1}$$

Donde:

- $h(t \mid X)$: función de riesgo condicional en el tiempo t dado el vector de covariables X,
- $h_0(t)$: función de riesgo basal, idéntica para todos los individuos,
- β: vector de coeficientes que cuantifican el efecto de cada covariable.

Este modelo se clasifica como *semiparamétrico* debido a que:

- Tiene una parte paramétrica, $e^{\beta^T X}$, dependiente de un número finito de parámetros β.
- Y una parte no paramétrica, $h_0(t)$, la cual no se especifica con una forma funcional fija, sino que se estima directamente de los datos [6].

Interpretación de los coeficientes β_j:

- Si $\beta_j > 0$: la covariable X_j incrementa el riesgo.
- Si $\beta_j < 0$: la covariable X_j disminuye el riesgo.

5.1.2. Estimación de parámetros mediante verosimilitud parcial

Dado que no se especifica una forma funcional para $h_0(t)$, se utiliza la *verosimilitud parcial*, la cual permite estimar los coeficientes β sin necesidad de conocer $h_0(t)$ [5].

Supongamos que se observan n individuos y que ocurren D eventos en los tiempos $t_1 < t_2 < \ldots < t_D$, con covariables X_i, y conjunto de riesgo $R(t_i)$.

Entonces:

La verosimilitud parcial se define como:

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)} \tag{5.2}$$

Y su log-verosimilitud correspondiente es:

$$\ell(\beta) = \sum_{i=1}^{D} \left[\beta^T X_i - \log \left(\sum_{j \in R(t_i)} \exp(\beta^T X_j) \right) \right] \tag{5.3}$$

La estimación de β se obtiene maximizando esta función:

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta) \tag{5.4}$$

5.1.3. Estimación del riesgo acumulado y supervivencia

Una vez estimado $\hat{\beta}$, la función de riesgo acumulado basal se estima mediante el *estimador de Breslow*, que es una generalización del estimador de Nelson-Aalen [10, 8]:

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{j \in R(t_i)} \exp(\hat{\beta}^T X_j)} \tag{5.5}$$

Luego, el riesgo acumulado individual se estima como:

$$\hat{H}(t \mid X) = \hat{H}_0(t) \cdot \exp(\hat{\beta}^T X) \tag{5.6}$$

Finalmente, la función de supervivencia condicional para un individuo con covariables X se estima mediante:

$$\hat{S}(t \mid X) = \exp\left(-\hat{H}_0(t) \cdot \exp(\hat{\beta}^T X)\right) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\beta}^T X)} \tag{5.7}$$

Donde:

$$\hat{S}_0(t) = \exp\left(-\hat{H}_0(t)\right) \tag{5.8}$$

es la función de supervivencia basal estimada.

Esta formulación proporciona una base sólida para modelar datos de tiempo hasta el evento, con flexibilidad suficiente para adaptarse a estructuras complejas y sin imponer supuestos estrictos sobre la forma del riesgo basal [5, 8]. Sin embargo, no considera una fracción no susceptible por sí solo y recordemos que asume la proporcionalidad en el riesgo.

En los capítulos siguientes, este marco servirá como punto de partida para extenderse hacia *modelos de cura* [15, 4].

Capítulo 6

Modelos de Curación

El modelo de curación o también llamado modelo de cura, es una extensión del análisis de supervivencia clásico que permite distinguir entre dos subpoblaciones de individuos dentro de un estudio: aquellos que son *susceptibles*, que eventualmente podrían experimentar el evento de interés y aquellos considerados *inmunes, curados o no susceptibles*, es decir, que nunca lo experimentarán, incluso si se los observa durante un tiempo indefinido [1, 11].

El nombre de este modelo proviene de su aplicación principal en el campo de la medicina, surge a partir de aplicaciones clínicas donde se observa que una fracción de los pacientes nunca presenta recurrencia, aún luego de un seguimiento prolongado, lo cual contradice el supuesto tradicional de los modelos clásicos, que asumen que todos los individuos experimentarán el evento si el seguimiento es lo suficientemente largo [5, 11]. Se utiliza para estudiar, por ejemplo, la recaída de enfermedades como el cáncer [15]. En estos casos, es habitual que una fracción de los pacientes no recaigan nunca, aun cuando hayan sido observados durante largos períodos, lo que contradice la hipótesis clásica del análisis de supervivencia, la cual asume que todos los individuos eventualmente experimentarán el evento [5, 11].

Otros ejemplos se pueden encontrar en finanzas (tiempo hasta que un banco quiebra), marketing (tiempo hasta que alguien compra un producto nuevo), seguros (tiempo hasta que se presenta una reclamación de garantía), demografía (tiempo hasta que alguien se casa), sociología (tiempo hasta la reincidencia de los presos liberados) y educación (tiempo necesario para resolver un problema), entre otros [1].

La literatura moderna en modelos de curación ha sido ampliamente desarrollada en trabajos recientes como los de Amico y Van Keilegom [1], quienes establecen un marco teórico general, y la obra de Peng y Yu [16], donde se describen en profundidad métodos, aplicaciones e implementación en software estadístico moderno.

6.1. Supuestos del modelo de curación

En medicina, el objetivo suele ser analizar el tiempo de recurrencia de una enfermedad, siendo la variable de interés el tiempo transcurrido T hasta que ocurre el evento. En el contexto de los modelos de cura, se reconoce que, cuando $t \to \infty$, una parte de los individuos no habrá experimentado el evento [1].

En estos casos, la función de distribución acumulada del tiempo hasta el evento $F(t) = P(T \leq t)$ no tiende a 1, sino a un valor menor:

$$\lim_{t\to\infty} F(t) = p < 1$$

Por tanto, la función de supervivencia $S(t) = P(T > t)$ no tiende a cero, como en los modelos clásicos, sino que se estabiliza en un valor positivo:

$$\lim_{t\to\infty} S(t) = 1 - p > 0$$

Este valor $1 - p$ se interpreta como la fracción de individuos que nunca experimentarán el evento, también conocida como fracción de supervivencia a largo plazo o *fracción no susceptible*, es decir, $S(\infty) > 0$.

6.2. Enfoques principales

Existen dos grandes enfoques para modelar esta fracción curada: el *modelo de mezcla* (mixture cure model), propuesto inicialmente por Boag (1949), y el *modelo de tiempo de promoción* (promotion time cure model), desarrollado posteriormente y explicado en detalle por Farewell [11].

El primero asume que la población está compuesta por dos grupos distintos: los curados, que nunca presentarán el evento (por ejemplo, una recaída), y los susceptibles, que eventualmente sí lo harán. Este modelo permite estimar de forma separada la probabilidad de pertenecer a cada grupo y el tiempo hasta que ocurre el evento entre los susceptibles.

En cambio, los modelos de tiempo de promoción no dividen explícitamente a la población, sino que suponen que el evento ocurre solo si se activan ciertos procesos biológicos o riesgos latentes. Si estos procesos no se activan nunca, el evento no ocurre, lo que da lugar de forma natural a una fracción curada.

En este trabajo se utilizará el primer enfoque, el modelo de mezcla.

6.3. Modelo de Curación de Mezcla

Este modelo consta de dos componentes clave, cada una asociada a una parte del proceso:

Componente de Incidencia

La parte de incidencia modela la probabilidad de que un individuo sea susceptible al evento, es decir, que eventualmente pueda experimentar dicho evento. Esta probabilidad usualmente se modela mediante una regresión logística:

$$\pi(X_i) = P(c_i = 1 \mid X_i) = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)} \tag{6.1}$$

Donde:

- $c_i \in \{0, 1\}$: si la variable es 1 indica que el individuo es susceptible, si es 0 no tenemos la información suficiente para decir que es susceptible.
- X_i: es un vector de covariables que afectan la probabilidad de ser susceptible.
- β: es el vector de parámetros asociados a la incidencia.

Componente de Latencia

Condicionada a que el individuo sea susceptible $c_i = 1$, se modela el tiempo hasta el evento mediante una distribución de supervivencia paramétrica:

$$S_{\text{laten}}(t \mid Z_i) = P(T_i > t \mid Z_i, c_i = 1)$$

Donde Z_i representa las covariables que afectan el tiempo hasta el evento.

Función de densidad de probabilidad

La función de densidad de probabilidad del tiempo hasta el evento, condicionada a que el individuo sea susceptible se deriva de la función de supervivencia para los susceptibles $S_{\text{laten}}(t \mid Z_i)$ mediante:

$$f_{\text{laten}}(t \mid Z_i) = -\frac{d}{dt} S_{\text{laten}}(t \mid Z_i) \tag{6.2}$$

Su forma específica depende de la distribución que se elija para modelar la latencia. En este trabajo se utiliza la distribución Weibull, ampliamente recomendada en modelos de curación por su flexibilidad y su interpretación clínica [16, 1].

6.3.1. Función de supervivencia poblacional

La función de supervivencia poblacional es una mezcla entre curados y susceptibles:

$$S_{\text{Total}}(t \mid X_i, Z_i) = 1 - \pi(X_i) + \pi(X_i) S_{\text{laten}}(t \mid Z_i) \tag{6.3}$$

donde:

- $1 - \pi(X_i)$: Fracción de individuos curados cuya supervivencia es 1.
- $\pi(X_i)$ Fracción de individuos susceptibles con supervivencia $S_{\text{laten}}(t \mid Z_i)$.

6.3.2. Estimación mediante Máxima Verosimilitud

Recordemos que la función de verosimilitud es una herramienta matemática que nos permite estimar los parámetros del modelo [8]. En los modelos de curación de mezcla, necesitamos considerar dos situaciones: si el evento fue observado o fue censurado [16].

Sea: t_i el tiempo observado para el individuo i

$$\delta_i = \begin{cases} 1 & \text{si el evento fue observado} \\ 0 & \text{si el evento fue censurado} \end{cases} \tag{6.4}$$

La verosimilitud del modelo para una muestra de tamaño n se escribe como:

$$L(\beta, \theta) = \prod_{i=1}^{n} \left[\pi(X_i) f_{\text{laten}}(t_i \mid Z_i)\right]^{\delta_i} \left[1 - \pi(X_i) + \pi(X_i) S_{\text{laten}}(t_i \mid Z_i)\right]^{1-\delta_i} \tag{6.5}$$

Donde:

- $f_{\text{laten}}(t_i \mid Z_i)$ es la densidad condicional del tiempo al evento para los susceptibles.
- θ representa los parámetros de la parte de latencia.

Es decir:

- Si $\delta_i = 1$, el individuo tuvo el evento: sabemos que es susceptible, y la contribución a la verosimilitud es la densidad de su tiempo de evento.
- Si $\delta_i = 0$, el individuo fue censurado: no sabemos si fue curado o no. Por eso, se considera una mezcla entre la probabilidad de estar curado y la de ser susceptible pero no haber tenido el evento aún.

En este proyecto utilizamos la estimación de parámetros por Máxima Verosimilitud ya que proporciona una estimación conjunta, eficiente y completa en una sola optimización.

6.4. Estimación mediante el algoritmo EM

Dado que c_i es una variable latente, es decir, no se observa directamente en el modelo, la verosimilitud está incompleta y no podemos maximizarla directamente como en los modelos clásicos. Para solucionar esto se utiliza el algoritmo de Expectation-Maximization (EM) y así estimar los parámetros del modelo. Este algoritmo implica un procedimiento iterativo, es decir, que se repite, y consiste en dos pasos:

6.4.1. Paso E (Expectation - Esperanza)

Calculamos la probabilidad esperada de que cada individuo sea susceptible ($c_i = 1$) dado lo que sí sabemos (su tiempo t_i y si fue o no censurado δ_i) y los parámetros actuales del modelo:

$$w_i^{(m)} = \mathbb{E}[c_i \mid t_i, \delta_i, \beta^{(m)}, \theta^{(m)}]$$

Si el individuo tuvo el evento: entonces $c_i = 1$, entonces $w_i^{(m)} = 1$. Si fue censurado: el modelo calcula la probabilidad de que sea susceptible y ese resultado será $w_i^{(m)}$.

Es así como este paso transforma la verosimilitud incompleta a una verosimilitud esperada.

6.4.2. Paso M (Maximization – Maximización)

Una vez obtenida la verosimilitud esperada, procedemos a maximizar dicha función para encontrar nuevos valores de los parámetros de incidencia β y de latencia θ.

En este paso se hace una regresión logística para los $w_i^{(m)}$ usando las covariables de incidencia y una estimación de modelo de supervivencia para los tiempos al evento, ponderados por $w_i^{(m)}$, entre los susceptibles.

El algoritmo EM se repite hasta que los parámetros convergen, es decir, hasta que los cambios entre iteraciones sucesivas son suficientemente pequeños.

El algoritmo EM presenta varias desventajas importantes. En primer lugar, su convergencia puede ser lenta, especialmente cuando la proporción de datos incompletos es elevada, no garantiza alcanzar el máximo global de la verosimilitud, por lo que es sensible a los valores iniciales y puede converger a máximos locales. Su desempeño puede verse afectado ante censura extrema o estructuras de datos poco informativas.

6.5. Implementación en R

Existen varios paquetes para ajustar modelos de curación:

- `smcure`: modelo semiparamétrico. (Cox)
- `flexsurvcure`: modelos paramétricos (Weibull, log-normal, etc.).
- `nltm`: modelos no lineales de tiempo.
- `cureit`: modelos semiparamétricos flexibles.
- `mixcure`: modelos de mezcla paramétricos.

En este trabajo se implementó una estimación manual mediante `nlm`, maximizando directamente la función de verosimilitud para un modelo de curación weibull [4, 9, 16, 17].

En contextos clínicos, como el estudio de la recaída por cáncer de mama, puede observarse que la función de riesgo no sigue un patrón constante ni estrictamente proporcional en el tiempo. Por ejemplo, el riesgo de recaída puede ser elevado en los primeros meses tras el tratamiento y disminuir gradualmente, o bien presentar formas más complejas dependiendo de factores clínicos y biológicos. Por ello, en los capítulos siguientes se utilizará una formulación paramétrica más flexible basada en la distribución Weibull, la cual permite capturar mejor estos comportamientos variables del riesgo a lo largo del tiempo. Esta elección es especialmente apropiada en el contexto de los modelos de cura, donde se reconoce que una fracción de los individuos podría considerarse curada, es decir, con probabilidad nula de experimentar el evento en el largo plazo [18, 5, 4, 15].

Una vez que hemos estimado los parámetros del modelo mediante máxima verosimilitud, es necesario evaluar su ajuste a los datos. Para ello, se pueden emplear herramientas diagnósticas adecuadas para este tipo de modelos, como los *residuos cuantílicos aleatorizados* (RQR, por sus siglas en inglés). Estos residuos permiten evaluar si el modelo describe adecuadamente la distribución observada de los tiempos al evento, incluso en presencia de censura [7]. En el próximo capítulo se detallará su definición, interpretación y utilidad práctica en el contexto de los modelos de cura.

Capítulo 7

Residuos Cuantílicos Aleatorizados (RQR)

En estadística, un residuo representa la discrepancia entre el valor observado y el valor predicho por un modelo. Sin embargo, en modelos complejos como los modelos de supervivencia, donde los datos pueden estar censurados o incluir una fracción de pacientes curados, los residuos tradicionales no son adecuados [9]. En estos casos, se recurre al uso de residuos transformados que tienen propiedades conocidas si el modelo es correctamente especificado.

Los residuos cuantílicos aleatorizados (Randomized Quantile Residuals, RQR) fueron propuestos originalmente por Dunn y Smyth [7] como una herramienta para diagnosticar la adecuación de modelos estadísticos, especialmente cuando la variable respuesta no sigue una distribución normal e incluso cuando la estructura incluye una mezcla de individuos susceptibles y curados, como ocurre en los modelos de curación. Su principal ventaja es que, si el modelo es correcto, estos residuos deberían seguir aproximadamente una distribución normal estándar, es decir $RQR \sim N(0,1)$. Esto permite evaluar visual y estadísticamente la calidad del ajuste del modelo mediante histogramas, gráficos Q-Q o pruebas de normalidad.

En esencia, los RQR se construyen a partir de la transformación de la función de distribución acumulada (CDF) del modelo ajustado y si el modelo está correctamente especificado, los valores observados transformados mediante la inversa de la función de distribución normal estándar deberían distribuirse aproximadamente como una normal estándar $N(0,1)$.

Para el cálculo de los RQR se utilizaron funciones del lenguaje R, pertenecientes al paquete `stats`, el cual se carga de manera predeterminada en cualquier sesión de R. En particular, se emplearon las funciones `runif()` para generar los valores aleatorios utilizados en la aleatorización necesaria en observaciones censuradas, `qnorm()` para transformar las probabilidades en residuos con distribución aproximadamente normal estándar, así como `pweibull()` y `dweibull()` para evaluar la supervivencia y la densidad Weibull en la parte de latencia del modelo ajustado.

7.1. Definición de los RQR

7.1.1. Caso no censurado

Para *observaciones no censuradas* (es decir, cuando se conoce el tiempo exacto del evento), el residuo RQR se define matemáticamente como:

$$r_i = \Phi^{-1}(\hat{F}_i(t_i)) \tag{7.1}$$

Donde:

- t_i : es el tiempo observado para el individuo i
- $\hat{F}_i(t_i)$: es la función de distribución acumulada ajustada por el modelo en t_i
- Φ^{-1} : es la inversa de la función de distribución acumulada de una normal estándar.

Esto equivale a tomar el valor observado, calcular su probabilidad acumulada según el modelo, y luego transformarlo a la escala normal estándar. Si el modelo es adecuado, los residuos r_i deberían comportarse como una muestra de $N(0, 1)$. No obstante, esta definición solo es válida cuando los datos no están censurados.

7.1.2. Caso censurado

En presencia de censura (cuando se desconoce el tiempo exacto del evento), es necesario introducir aleatorización para que los residuos mantengan su propiedad asintótica de normalidad bajo el modelo correcto.

En tal caso, el residuo RQR se define como:

$$r_i = \Phi^{-1}(u_i) \quad \text{con } u_i \sim \text{Uniforme}(\hat{F}_i(t_i), 1)$$

O de manera equivalente:

$$r_i = \Phi^{-1}(1 - \hat{S}_i^{\text{TOTAL}}(t_i) \cdot v_i) \quad \text{con } v_i \sim \text{Uniforme}(0, 1) \tag{7.2}$$

Donde:

- $\hat{S}_i^{\text{TOTAL}}(t_i) = 1 - \hat{F}_i(t_i)$: representa la supervivencia marginal ajustada por el modelo en t_i
- v_i : es una variable aleatoria que introduce variabilidad en el residuo para los datos censurados.

Esta aleatorización asegura que los residuos censurados también sigan una distribución aproximadamente normal estándar, siempre que el modelo esté correctamente especificado.

Logrando así una herramienta diagnóstica coherente tanto para observaciones censuradas como no censuradas.

7.2. Definición de los RQR en Modelos de Cura

Sea T_i el tiempo observado para el individuo i, con indicador de evento δ_i

$$\delta_i = \begin{cases} 1 & \text{si el evento fue observado} \\ 0 & \text{si el evento fue censurado} \end{cases}$$

En los modelos de curación de mezcla, usualmente la probabilidad de que un individuo sea susceptible se modela mediante:

$$\pi_i = \text{logit}^{-1}(X_i^\top \widehat{\beta}),$$

mientras que, condicionando a que el individuo sea susceptible, la supervivencia del tiempo al evento viene dada por:

$$S_i(t) = S_0(t)^{\exp(\psi_i)},$$

donde $S_0(t)$ para nuestro estudio es la supervivencia Weibull paramétrica del modelo de latencia, y $\psi_i = Z_i^\top \widehat{\gamma}$ corresponde a los efectos de las covariables en la latencia.

A partir de estos elementos, la función de distribución acumulada total del tiempo al evento en un modelo de cura es:

$$F_i(t) = (1 - \pi_i) + \pi_i \left[1 - S_i(t)\right].$$

Caso 1: Evento observado ($\delta_i = 1$)

Cuando el evento ocurre en t_i, el residuo RQR se define como:

$$\text{RQR}_i = \Phi^{-1}(F_i(t_i)) = \Phi^{-1}(\pi_i \left[1 - S_i(t_i)\right])$$

Caso 2: Observación censurada ($\delta_i = 0$)

En un modelo de cura la aleatorización se implementa mediante:

$$U_i = 1 - \left[(1 - \pi_i)v_i + \pi_i S_i(t_i)\right], \qquad v_i \sim U(0, 1),$$

y el residuo se define como:

$$\text{RQR}_i = \Phi^{-1}(U_i).$$

Este mecanismo garantiza que, bajo un modelo correctamente especificado, los RQR se distribuyan aproximadamente como una $N(0, 1)$, independientemente de la presencia de censura o mezcla.

7.3. Interpretación

Si el modelo está bien ajustado, los residuos RQR deberían distribuirse aproximadamente como una normal estándar $N(0,1)$. Por tanto:

- Un histograma de los residuos debería aproximarse a la campana de Gauss
- Un gráfico Q-Q debería alinearse con la diagonal
- Las pruebas de normalidad no deberían rechazar la hipótesis nula de normalidad

Grandes desviaciones de esta distribución en los residuos indican problemas de especificación del modelo, mala elección de la función de supervivencia base o falta de ajuste de los efectos de las covariables.

Para el cálculo de los residuos cuánticos aleatorizados (RQR) se utilizaron exclusivamente funciones base del lenguaje `R`, pertenecientes al paquete `stats`. En particular, se emplearon las funciones `runif()` para generar los valores aleatorios utilizados en la aleatorización necesaria en observaciones censuradas, `qnorm()` para transformar las probabilidades en residuos con distribución aproximadamente normal estándar, así como `pweibull()` y `dweibull()` para evaluar la supervivencia y la densidad Weibull en la parte de latencia del modelo ajustado.

Existen múltiples pruebas de normalidad disponibles en la literatura estadística, pero para la validación de los residuos en nuestro modelo de curación Weibull se seleccionó específicamente la prueba de Anderson-Darling sobre otras alternativas como Shapiro-Wilk, Kolmogorov-Smirnov, Cramér-von Mises y Jarque-Bera por su *superior sensibilidad en las colas de la distribución*, característica crítica para detectar desviaciones sutiles de normalidad en residuos RQR.

Mientras la prueba de Shapiro-Wilk pierde potencia con muestras grandes como la nuestra ($n = 286$), Kolmogorov-Smirnov subestima discrepancias en extremos y no es invariante bajo transformaciones. Por su parte, Cramér-von Mises asigna igual peso a todas las desviaciones y Jarque-Bera se limita a evaluar únicamente asimetría y curtosis, mostrando pobre rendimiento con tamaños de muestra moderados.

Anderson-Darling supera estas limitaciones mediante una función de peso $w(t) = [F(t)(1-F(t))]^{-1}$ que enfatiza precisamente las discrepancias en valores extremos, donde los problemas de especificación del modelo suelen manifestarse con mayor claridad.

7.4. Prueba de normalidad Anderson-Darling

Esta es una prueba estadística es ampliamente utilizada para evaluar si una muestra de datos proviene de una distribución específica, en este caso, una distribución normal estándar $N(0, 1)$.

7.4.1. Hipótesis de prueba

- H_0: Los residuos $r_i \sim N(0, 1)$ (el modelo está bien especificado).

- H_1: Los residuos no siguen una distribución normal estándar (el modelo podría estar mal especificado).

7.4.2. Estadístico de prueba

Dado un conjunto ordenado de residuos $r_1 \leq r_2 \leq \cdots \leq r_n$, y sea $F(r)$ la función de distribución acumulada de la normal estándar, el estadístico de Anderson-Darling se define como:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i - 1) \left(\ln F(r_i) + \ln(1 - F(r_{n+1-i})) \right) \right] \tag{7.3}$$

Este estadístico mide la distancia entre la distribución empírica de los residuos y la distribución normal estándar. Cuanto mayor sea A^2, mayor es la evidencia en contra de la normalidad.

7.4.3. Interpretación

- Si el valor A^2 es pequeño y su p-valor asociado es mayor que el nivel de significancia, por ejemplo 0.05, no se rechaza H_0; es decir, no hay evidencia de que los residuos se desvíen de la normalidad.

- Si el p-valor es pequeño, por ejemplo, menor a 0.05, se rechaza H_0, lo que sugiere que el modelo puede no estar bien ajustado.

Capítulo 8

Aplicación

8.1. Planteamiento

El cáncer de mama representa una de las principales causas de mortalidad oncológica a nivel global, siendo responsable del 15.7 % de las muertes por cáncer en las Américas [14]. Esta enfermedad se caracteriza por el crecimiento descontrolado de células malignas en el tejido mamario, con capacidad para metastatizar a otros órganos. Su desarrollo está influenciado por múltiples factores, incluyendo características hormonales (como la expresión de receptores de estrógeno y progesterona), alteraciones genéticas y factores ambientales.

Los biomarcadores moleculares, definidos como características biológicas medibles indican procesos fisiológicos, estados patogénicos o respuestas terapéuticas, han revolucionado la oncología moderna. Permiten estratificar pacientes, predecir evolución clínica y personalizar tratamientos entre ellos, el biomarcador *ESR1 (Receptor de Estrógeno Alfa)* constituye un marcador pronóstico y predictivo ampliamente validado en cáncer de mama [2, 19].

Este capítulo tiene como objetivo *aplicar un modelo de cura de mezcla con distribución Weibull* a un conjunto de datos reales provenientes de pacientes con cáncer de mama en estadio temprano [12, 13]. Se busca evaluar la relación entre la expresión génica del biomarcador *ESR1* y el riesgo de recaída, utilizando datos del estudio GSE2034 [13]. Esta aplicación tiene como finalidad ilustrar la implementación práctica del modelo desarrollado, validar su ajuste mediante herramientas diagnósticas como los residuos cuantílicos aleatorizados (RQR) [7] y explorar su utilidad para la estratificación del riesgo. El análisis permite demostrar el potencial de los modelos de cura como herramienta de apoyo a decisiones clínicas basadas en biomarcadores moleculares [15].

8.2. Descripción de los datos

Para este estudio se utilizó la base de datos pública GSE2034, disponible en el repositorio Gene Expression Omnibus (GEO), que contiene datos de expresión génica y clínica de 286 pacientes con cáncer de mama sin afectación ganglionar y sin tratamiento adyuvante [13]. La expresión génica fue medida mediante la plataforma Affymetrix HG-U133A, que permite cuantificar la expresión de más de 20,000 genes por muestra [12]. Se seleccionó la sonda `205225_at`, que mide la *expresión del gen ESR1 (Receptor de Estrógeno Alfa).*

Este gen es un biomarcador ampliamente validado, ya que su expresión se ha asociado con la respuesta a terapia hormonal y con el pronóstico clínico. La sonda fue empleada en el estudio de Wang et al. [19] como parte de un panel de 76 genes predictivos de metástasis, y también fue utilizada por Budczies et al. [3] para determinar puntos de corte óptimos mediante la herramienta *Cutoff Finder*, encontrando valores (escala $\log_2$) como umbral para clasificar tumores. Por ello, el marcador *ESR1* es un componente clave para la estratificación del riesgo.

En particular se tomó el subconjunto de datos descritos en la Tabla 8.1:

Tabla 8.1: Descripción de variables utilizadas en el modelo de cura Weibull

Variable	**Descripción**	**Relevancia en el modelo**
`marker`	Expresión $\log_2$ del gen ESR1 (sonda 205225_at).	Biomarcador principal para estimar riesgo de recaída [19].
`time`	Tiempo en meses desde el diagnóstico hasta recaída o último seguimiento.	Variable temporal fundamental para el análisis de supervivencia.
`status`	Indicador de estado: 1 = recaída (evento observado), 0 = censura.	Diferencia entre pacientes que experimentaron el evento y los que no.

8.3. Metodología y Desarrollo de los Modelos de Supervivencia

Con el objetivo de evaluar el efecto del biomarcador *ESR1* sobre el riesgo y el tiempo hasta la recaída, se ajustaron dos modelos de cura de mezcla con distribución Weibull: un modelo nulo, sin covariables, y un modelo completo, que incluye transformaciones del marcador como covariables tanto en la parte de incidencia como en la de latencia. Esta comparación permite determinar si el biomarcador tiene un efecto significativo en alguno de los componentes del modelo.

8.4. Modelo de Curación de Mezcla Nulo (Sin Covariables)

El modelo nulo de curación Weibull es una forma básica del modelo de mezcla de curación que no incluye covariables. Su objetivo es estimar la proporción de pacientes considerados curados y modelar el tiempo hasta la recaída entre los susceptibles usando una distribución Weibull [11, 15]. Este modelo sirve como referencia para evaluar si la inclusión de covariables mejora el ajuste y la interpretación clínica.

8.4.1. Formulación del Modelo Nulo de Curación

La función de supervivencia total bajo el modelo de curación nulo es:

$$S(t) = (1 - \pi_0) + \pi_0 \cdot \exp\left[-\left(\frac{t}{\lambda}\right)^{\alpha}\right], \tag{8.1}$$

donde:

- π_0 es la proporción de pacientes susceptibles (probabilidad de recaer eventualmente, complemento de la fracción curada).
- λ es el parámetro de escala de la distribución Weibull.
- α es el parámetro de forma de la distribución Weibull.

La función de verosimilitud del modelo de curación nulo es:

$$L(\pi_0, \lambda, \alpha) = \prod_{i=1}^{n} \left[\pi_0 f(t_i)\right]^{c_i} \left[(1 - \pi_0) + \pi_0 S(t_i)\right]^{1-c_i} \tag{8.2}$$

Y la función de log-verosimilitud para datos de supervivencia con censura se expresa como:

$$\ell(\theta) = \sum_{i=1}^{n} \left[c_i \log\left(\pi_0 f(t_i)\right) + (1 - c_i) \log\left((1 - \pi_0) + \pi_0 S(t_i)\right)\right], \tag{8.3}$$

- $\theta = (\pi_0, \lambda, \alpha)$: Parámetros del modelo
- c_i: Indicador de evento ($c_i = 1$ si recayó, $c_i = 0$ si censurado)
- $f(t_i)$: Función de densidad Weibull para el tiempo t_i
- $S(t_i)$: Función de supervivencia Weibull para el tiempo t_i

Teniendo en cuenta que la función de densidad Weibull es:

$$f(t) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left[-\left(\frac{t}{\lambda}\right)^{\alpha}\right] \tag{8.4}$$

Y la función de supervivencia Weibull es:

$$S(t) = \exp\left[-\left(\frac{t}{\lambda}\right)^{\alpha}\right] \tag{8.5}$$

8.4.2. Estimación en R

La estimación se realizó mediante máxima verosimilitud utilizando una función personalizada:

```
#-------función de verosimilitud--------------
ll.cure.wei <- function(parameters, t.i, c.i){
  p <- plogis(parameters[1])
  lambda <- exp(parameters[2])
  alpha <- exp(parameters[3])
  f.i <- dweibull(t.i, shape=alpha, scale=lambda)
  S.i <- 1 - pweibull(t.i, shape=alpha, scale=lambda)
  -sum(c.i * log(p * f.i) + (1 - c.i) * log((1 - p) + p * S.i))
}
```

Se estimaron los parámetros con la función `nlm() (Non-Linear Minimization)` del paquete stats:

```
#------(optimizando) Maximizando funcion------
mle.cure.wei <- nlm(ll.cure.wei, c(0, 2, 0),
                    t.i = db_205225_at$time,
                    c.i = db_205225_at$status,
                    hessian = TRUE)
```

8.4.3. Resultados del Modelo Nulo de Curación

Tabla 8.2: Estimación de parámetros del modelo nulo de cura Weibull

Parámetro	Estimación	Error estándar	Estadístico z	Valor p
Proporción susceptible (logit)	-0.490	0.124	-3.96	$<$0.001
Escala Weibull (log λ)	3.598	0.065	55.12	$<$0.001
Forma Weibull (log α)	0.497	0.082	6.10	$<$0.001

Interpretación Clínica

- **Fracción curada**: $1 - \hat{\pi}_0 = 62\,\%$, lo que indica que aproximadamente 6 de cada 10 pacientes no experimentan recaída a largo plazo.
- **Parámetro de forma** ($\hat{\alpha} = 1.64$): Sugerente de un riesgo de recaída creciente con el tiempo entre las pacientes susceptibles.
- **Parámetro de escala** ($\hat{\lambda} = 36.55$ meses): Sugiere que las recaídas, entre quienes son susceptibles, tienden a ocurrir al rededor de los primeros 36 meses.

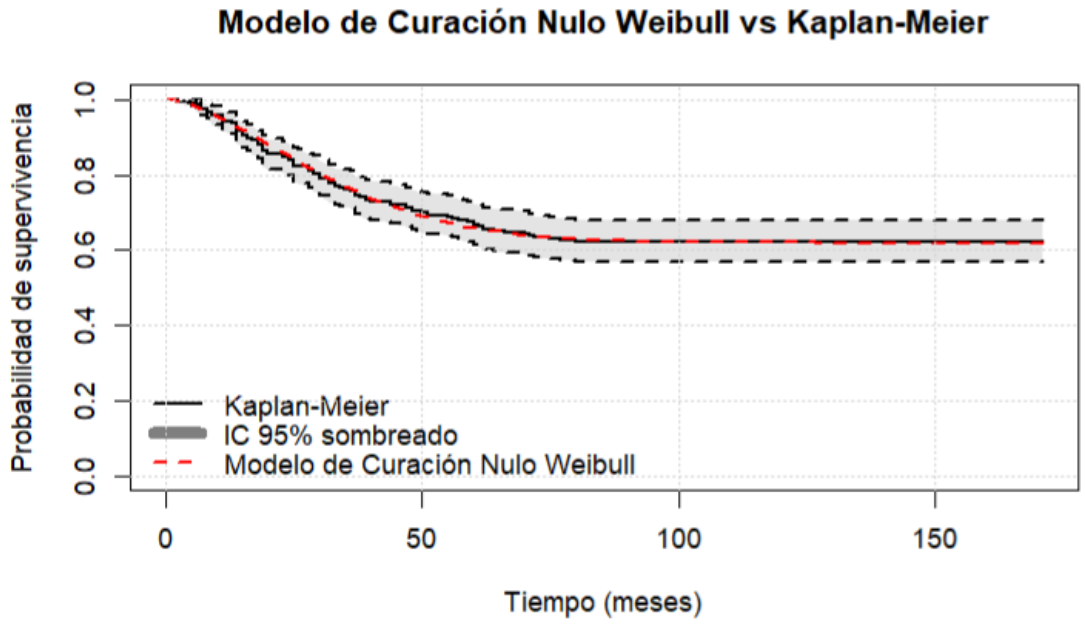


Figura 8.1: Comparación entre la curva de Kaplan-Meier y el modelo de Nulo de Curación Weibull ajustado.

Los resultados obtenidos hasta este punto muestran que tanto la regresión logística utilizada para modelar la fracción susceptible como la distribución Weibull empleada para la componente de latencia se ajustan adecuadamente a los datos, proporcionando estimaciones coherentes con el comportamiento observado en la curva de Kaplan–Meier. Habiendo verificado que el modelo de curación nulo representa de manera satisfactoria la estructura básica de los tiempos hasta la recaída, el siguiente paso consiste en evaluar cómo varía el riesgo cuando se incorpora la expresión génica del biomarcador ESR1, que hemos denotado como marker, con el objetivo de determinar si este marcador modula tanto la probabilidad de pertenecer al grupo susceptible como la dinámica temporal del riesgo entre quienes pueden recaer.

8.5. Modelo de Curación de Mezcla con Covariables

8.5.1. Selección de la forma funcional del marcador ESR1

Antes de definir el modelo de mezcla de cura bajo distribución Weibull, se llevó a cabo un análisis exploratorio para identificar la forma funcional más adecuada del biomarcador *ESR1* en ambas componentes del modelo (incidencia y latencia).

Para la *incidencia*, inicialmente se evaluaron transformaciones polinomiales con el objetivo de capturar posibles curvaturas en la relación entre el marcador y la probabilidad de ser susceptible. Sin embargo, los términos polinomiales no resultaron significativos y no lograron representar adecuadamente el patrón observado en los datos.

A partir de la evidencia gráfica y de la hipótesis clínica de que podría existir un punto donde la pendiente del efecto del marcador cambia, se exploraron transformaciones por tramos mediante funciones indicadoras del tipo:

$$x_C^* = x \cdot I(x < C),$$

generando una sucesión de posibles puntos de corte C sobre el rango del marcador. Entre estos valores, sólo un punto produjo una pendiente estadísticamente significativa en el componente de incidencia. El valor óptimo encontrado fue $C = 12.9$, cuya transformación generó un coeficiente significativo ($p \approx 0.049$). Este comportamiento sugiere un posible efecto umbral, consistente con enfoques como *Cutoff Finder* [3].

Por ello, la transformación escalonada

$$x^* = x \cdot I(x < 12.9)$$

fue seleccionada para el modelo final en la componente de incidencia.

Por otro lado en el *componente de latencia*, el objetivo no era identificar un punto de cambio sino modelar adecuadamente la velocidad del riesgo entre las pacientes susceptibles. En este caso, el comportamiento del marcador mostró una tendencia aproximadamente lineal, por lo que se utilizó una transformación polinómica ortogonal de primer grado, resultando ser también muy significativa:

$$z(x) = \texttt{poly(marker,1)}.$$

El uso de polinomios ortogonales permite:

- evitar colinealidad
- y mejorar la estabilidad numérica en la estimación.

8.5.2. Modelo de mezcla de cura Weibull

Con el objetivo de modelar la evolución de recaídas en pacientes con cáncer de mama, se ajustó un modelo de cura paramétrico bajo la distribución Weibull, incorporando la expresión del biomarcador *ESR1* como covariable. Este tipo de modelo permite separar a las pacientes en dos grupos: aquellas que son susceptibles a recaer y aquellas que podría ser inmunes (es decir, que no recaerán en el largo plazo).

Este modelo se estructura en dos componentes principales:

- *Componente de incidencia*: modela la probabilidad de ser *susceptible* (recaer eventualmente), mediante una regresión logística $\pi(x)$.

  $$\pi(x) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 \cdot x^*)\right)},$$

  donde $x^* = x \cdot I(x < 12.9)$ es la transformación escalonada seleccionada.

- *Componente de latencia*: modela el tiempo hasta la recaída *entre las pacientes susceptibles*, usando un modelo Weibull.

  $$S_u(t \mid x) = \exp\left\{-\left(\frac{t}{\lambda}\right)^{\alpha} \cdot \exp(\gamma_0 + \gamma_1 \cdot z(x))\right\},$$

  donde:

  $\lambda > 0$ es el parámetro de escala,
  $\alpha > 0$ es el parámetro de forma,
  $z(x)$ es la transformación ortogonal aplicada al marcador.

Matemáticamente, la *función de supervivencia poblacional* (es decir, para todas las pacientes) se define como:

$$S(t \mid x) = 1 - \pi(x) + \pi(x) \cdot S_u(t \mid x),$$

donde:

- $\pi(x)$ es la probabilidad de ser susceptible (y por tanto recaer),
- $S_u(t \mid x)$ es la función de supervivencia Weibull de las pacientes susceptibles,
- x es una covariable basada en la expresión génica de `ESR1`.

8.5.3. Supervivencia poblacional

Por tanto, la supervivencia poblacional final es:

$$S(t \mid x) = (1 - \pi(x)) + \pi(x) \cdot \exp\left\{-\left(\frac{t}{\lambda}\right)^{\alpha} \cdot \exp(\gamma_1 \cdot z(x))\right\}$$

8.5.4. Función de verosimilitud

Sea t_i el tiempo observado para la paciente i, c_i su indicador de evento (1 si recayó, 0 si fue censurada), y x_i su marcador. La función de verosimilitud se escribe como:

$$L(\theta) = \prod_{i=1}^{n} \left[\pi(x_i) \cdot f_u(t_i \mid x_i)\right]^{c_i} \cdot \left[(1 - \pi(x_i)) + \pi(x_i) \cdot S_u(t_i \mid x_i)\right]^{1-c_i},$$

Por tanto, la log-verosimilitud es:

$$\ell(\theta) = \sum_{i=1}^{n} \left[c_i \cdot \log(\pi(x_i) \cdot f_u(t_i \mid x_i)) + (1 - c_i) \cdot \log(1 - \pi(x_i) + \pi(x_i) \cdot S_u(t_i \mid x_i))\right],$$

donde $f_u(t \mid x)$ es la densidad de Weibull ajustada entre los susceptibles.

Los parámetros del modelo $\theta = (\beta_0, \beta_1, \gamma, \log\lambda, \log\alpha)$ se estimaron por máxima verosimilitud mediante programación en `R`.

8.5.5. Implementación computacional

Para estimar los parámetros del modelo por máxima verosimilitud, se construyeron dos matrices de diseño separadas, una para cada componente del modelo, utilizando la función `model.matrix()` de `R`. Estas matrices fueron transformadas de manera distinta, de acuerdo con el tipo de relación esperada y la interpretación clínica:

- *Matriz para incidencia*:

  $$\texttt{mat.incid} \leftarrow \texttt{model.matrix}(\sim I(marker \cdot (marker < 12.9)))$$

  Esta transformación genera dos columnas:

  - Un intercepto (β_0),
  - Una variable escalonada $x^* = marker \cdot I(marker < 12.9)$, que es igual al marcador si este es menor que 12.9 y cero en otro caso.

  Esta transformación permite capturar un posible *efecto umbral*, consistente con hallazgos clínicos donde valores bajos de *ESR1* podrían estar asociados a mayor riesgo de recaída. El uso de puntos de corte empíricos se respalda con herramientas como *Cutoff Finder* [3], y en este caso se seleccionó 12.9 por presentar un estadístico $z \approx -1.97$ y un valor $p \approx 0.049$, lo que sugiere asociación estadísticamente significativa.

- *Matriz para latencia*:

$$\texttt{mat.laten} \leftarrow \texttt{model.matrix}(\sim \texttt{poly}(marker, 1))[, -1]$$

  Esta transformación aplica una función polinómica ortogonal de primer grado al marcador para evitar colinealidad entre términos. Se elimina el intercepto automático (la primera columna), ya que el modelo Weibull ya contiene parámetros de base (escala y forma), y añadir un intercepto adicional sería redundante. Esta práctica es común para mejorar la estabilidad numérica en modelos paramétricos [5].

La estimación de parámetros se realizó por máxima verosimilitud y la optimización se llevó a cabo con `nlm()`.

La implementación de la función de verosimilitud combinada se implementó como:

```
#--------------función de verosimilitud--------------------
ll.cure.wei_COMBINADO <- function(parameters, t.i, c.i,
X.incid, X.laten){
  # Incidencia
  eta_cov <- X.incid %*% parameters[1:ncol(X.incid)]
  p_cov <- plogis(eta_cov)

  # Latencia
  psi_cov <- X.laten %*% parameters[(ncol(X.incid)+1):
  (ncol(X.incid)+ncol(X.laten))]
  lambda <- exp(parameters[ncol(X.incid) + ncol(X.laten) + 1])
  alpha <- exp(parameters[ncol(X.incid) + ncol(X.laten) + 2])

  h0 <- dweibull(t.i, shape = alpha, scale = lambda)
  / (1 - pweibull(t.i, shape = alpha, scale = lambda))
  S0 <- 1 - pweibull(t.i, shape = alpha, scale = lambda)

  h.i <- h0 * exp(psi_cov)
  S.i <- S0^exp(psi_cov)
  f.i <- h.i * S.i

  -sum(c.i * log(p_cov * f.i) + (1 - c.i) * log((1 - p_cov)
  + p_cov * S.i))}

# Optimizacion
mle.cure.wei_COMBINADO <- nlm(ll.cure.wei_COMBINADO, p = initial_params,
                              t.i = db_205225_at$time,
                              c.i = db_205225_at$status,
                              X.incid = mat.incid,
                              X.laten = mat.laten,
                              hessian = TRUE)
```

8.5.6. Resultados del Modelo de Mezcla de Cura Weibull

Tabla 8.3: Estimaciones del modelo de cura Weibull con marcador ESR1

Parámetro	Estimación	Error estándar	z-valor	p-valor
Incidencia: intercepto	-0.240	0.177	-1.36	0.174
Incidencia: $x^* = x \cdot I(x < 12.9)$	-0.047	0.024	-1.97	0.049
Latencia: marcador (poly)	-7.825	1.775	-4.41	$<$0.001
$\log(\lambda)$	3.560	0.059	59.93	$<$0.001
$\log(\alpha)$	0.586	0.081	7.19	$<$0.001

Interpretación

Se ajustó un modelo de cura combinado con distribución Weibull a los datos, utilizando una función logística para modelar la probabilidad de pertenecer a la fracción susceptible (*incidencia*) y un modelo de Weibull para el tiempo hasta la recaída entre las pacientes susceptibles (*latencia*).

- El coeficiente de incidencia asociado a `I(marker*(marker <12.9))` fue negativo y estadísticamente significativo ($p = 0.049$), lo que indica que, a medida que aumenta el valor del marcador (solo para valores menores a 12.9), la probabilidad de ser susceptible disminuye, siendo más probable que el individuo pertenezca al grupo Çurado.º no susceptible.
- El coeficiente de latencia fue también negativo y altamente significativo ($\hat{\gamma}_1 = -7.83$, $p < 0.001$). Esto sugiere que este marcador tiene una fuerte influencia en el tiempo de supervivencia. Entre las pacientes susceptibles, una mayor expresión del marcador `ESR1` se asocia con un mayor tiempo de supervivencia y por tanto una reducción en el riesgo instantáneo de que ocurra una recaída entre los susceptibles.
- Dado que $\alpha > 1$, la función de riesgo es creciente en el tiempo, lo cual es consistente con enfermedades progresivas como el cáncer.
- $\lambda \approx 35 meses$ representa la escala temporal sobre la cual ocurren la mayoría de las recaídas.

Desde el punto de vista biológico, *ESR1* es un biomarcador ampliamente reconocido por su asociación con mejor pronóstico y buena respuesta a la terapia hormonal en cáncer de mama. Sin embargo, este marcador puede tener un comportamiento más complejo en ciertos contextos. En particular, cuando ocurre una mutación [2]. Sin embargo, no es nuestro caso.

8.6. Análisis de Incidencia: Factores que influyen en la Susceptibilidad

Tras el ajuste del modelo de mezcla de cura Weibull con expresión génica, se realizó un análisis detallado de la parte de *incidencia*, con el objetivo de representar la probabilidad de *ser susceptible* en función del nivel del marcador *ESR1*. Este análisis se basa en la misma especificación del modelo logístico utilizado previamente, empleando los coeficientes estimados.

Del ajuste del modelo se obtuvieron los siguientes coeficientes (Tabla 8.3):

$$\hat{\beta}_0 = -0.240, \qquad \hat{\beta}_1 = -0.047 \; (p = 0.049).$$

La interpretación es la siguiente:

- El intercepto $\widehat{\beta}_0$ representa el logit de la probabilidad de ser susceptible para un individuo con marcador igual a 0 (es el valor base del modelo).
- El coeficiente $\widehat{\beta}_1$, negativo y estadísticamente significativo, indica que para valores de `marker` menores a 12.9, un aumento en el marcador se asocia con menores *odds* de ser susceptible, aumentando así la probabilidad de encontrarse en la fracción curada.

Para cuantificar este efecto se calcularon los *odds ratios* (OR) y sus intervalos de confianza. El OR del marcador es:

$$OR_{\text{marker}} = 0.954, \qquad IC_{95\,\%} = [0.9103,\ 0.9998].$$

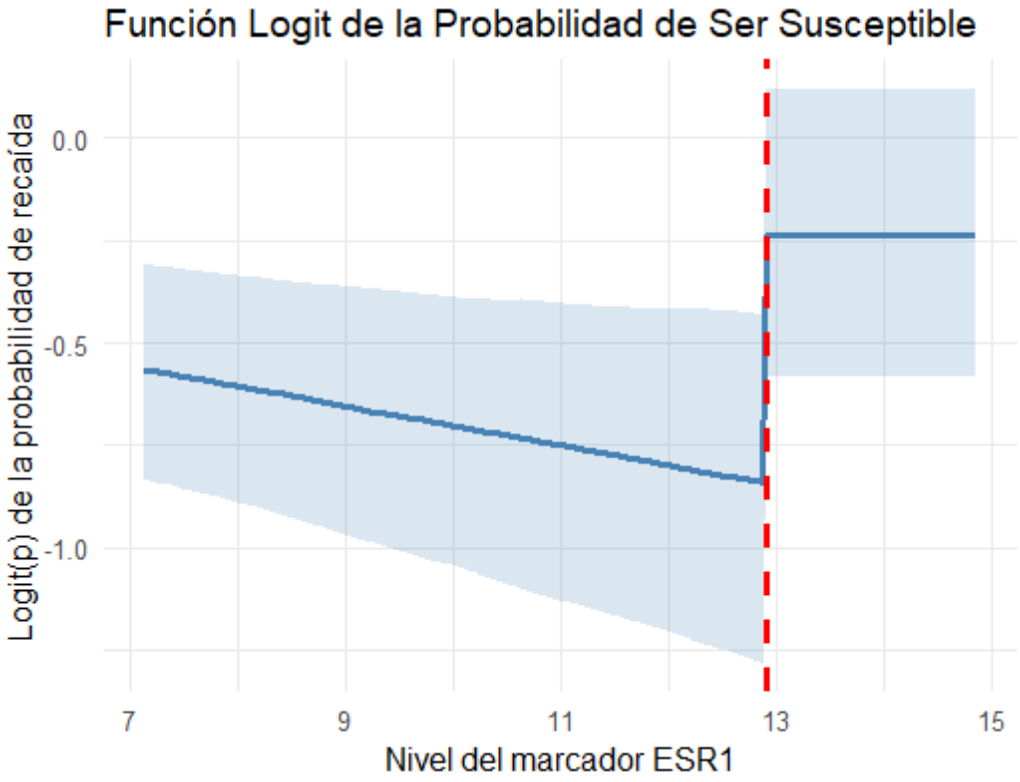
Dado que todo el intervalo se encuentra por debajo de 1, se confirma que el marcador ESR1 ejerce un efecto protector en términos de susceptibilidad, pero únicamente cuando sus valores son menores a 12.9. A partir de este punto, el efecto se estabiliza y el modelo no estima cambios adicionales en la incidencia.

La Figura 8.2 a) ilustra la función logit $\eta(\text{marker})$ con intervalos de confianza del 95 % obtenidos mediante simulación Monte Carlo. Se observa un cambio claro en la pendiente en el umbral *marker* = 12.9, punto a partir del cual el efecto del biomarcador sobre la probabilidad de ser susceptible deja de variar.

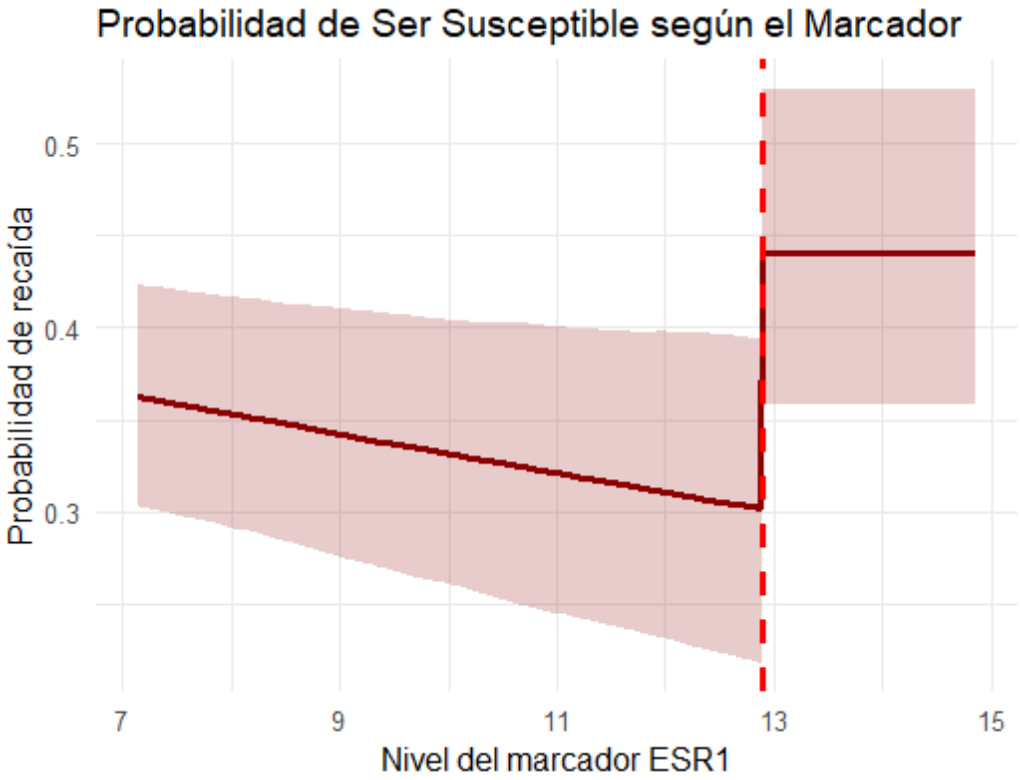
Posteriormente, se aplicó la función logística inversa para obtener la probabilidad directa de ser susceptible:

$$p(\text{marker}) = \frac{1}{1 + e^{-\eta(\text{marker})}}.$$

La Figura 8.2 b) muestra la probabilidad estimada de ser susceptible (no curarse) en función del marcador *ESR1*, confirmando el aplanamiento de la curva a partir del punto de corte. El procedimiento completo para el cálculo de intervalos de confianza mediante simulación Monte Carlo, se detalla en el Anexo correspondiente.



(a) Función logit η(marker) con IC del 95 % por Monte Carlo. Se observa un cambio en la pendiente, umbral $marker = 12.9$, a partir del cual el efecto del biomarcador sobre la probabilidad de ser susceptible deja de variar.



(b) Probabilidad estimada de ser susceptible en función del marcador *ESR1*. Los IC del 95 % reflejan la incertidumbre en la estimación.

Figura 8.2: Relación entre el marcador *ESR1* y la probabilidad de ser susceptible.

Algorithm 1 Simulación Monte Carlo y generación de las curvas η(marker) y p(marker)

procedure GRAFICAS INCIDENCIA($\hat{\beta}_0, \hat{\beta}_1, \widehat{\Sigma}$)
  Definir secuencia del marcador
  $marker \leftarrow [5, 5.1, \ldots, 20]$ ▷ Rango observado de ESR1
  Simular parámetros por Monte Carlo
  $B \leftarrow 5000$ ▷ Número de simulaciones
  for $b = 1$ to B **do**
    $(\beta_0^{(b)}, \beta_1^{(b)}) \leftarrow \text{MVN}((\hat{\beta}_0, \hat{\beta}_1), \widehat{\Sigma})$
    $\eta^{(b)}(marker) \leftarrow \beta_0^{(b)} + \beta_1^{(b)} \cdot marker$
    $p^{(b)}(marker) \leftarrow \frac{1}{1+\exp(-\eta^{(b)}(marker))}$
  end for
  Calcular bandas de confianza
  $\eta_{LCI}(marker),\ \eta_{UCI}(marker) \leftarrow 2.5\,\%, 97.5\,\%$ de $\eta^{(b)}(marker)$
  $p_{LCI}(marker),\ p_{UCI}(marker) \leftarrow 2.5\,\%, 97.5\,\%$ de $p^{(b)}(marker)$
  Identificar punto de corte
  $corte \leftarrow 12.9$
  Generar curvas
  Graficar $\eta(marker)$ con bandas de confianza
  Agregar línea punteada vertical en *corte*
  Graficar $p(marker)$ con bandas de confianza
  Agregar línea punteada vertical en *corte*
end procedure

Algorithm 2 Cálculo del Odds Ratio (OR) e intervalos de confianza del 95 %

procedure CALCULAROR($\hat{\beta}_0, \hat{\beta}_1, se(\hat{\beta}_0), se(\hat{\beta}_1)$)
  Calcular el OR del intercepto
  $OR_{\text{int}} \leftarrow \exp(\hat{\beta}_0)$
  Calcular el OR asociado al marcador
  $OR_{\text{marker}} \leftarrow \exp(\hat{\beta}_1)$
  Obtener límites del intervalo del 95 % para $\hat{\beta}_1$
  $\beta_{\text{inf}} \leftarrow \hat{\beta}_1 - 1.96 \cdot se(\hat{\beta}_1)$
  $\beta_{\text{sup}} \leftarrow \hat{\beta}_1 + 1.96 \cdot se(\hat{\beta}_1)$
  Convertir límites del intervalo a escala OR
  $IC_{\text{inf}} \leftarrow \exp(\beta_{\text{inf}})$
  $IC_{\text{sup}} \leftarrow \exp(\beta_{\text{sup}})$
  Salida: $OR_{\text{int}}, OR_{\text{marker}}, \left[IC_{\text{inf}}, IC_{\text{sup}}\right]$
end procedure

8.7. Análisis de Latencia y Visualización de Efectos Condicionales

Para ilustrar los efectos combinados del marcador *ESR1* se generaron representaciones visuales de las funciones de riesgo $h_i(t)$, supervivencia $S_i(t)$ y densidad $f_i(t)$ entre quienes son susceptibles para distintos valores del biomarcador.

Procedimiento para generar las curvas

Para construir las curvas, se utilizaron los parámetros estimados del modelo de cura Weibull combinado, en particular:

- El coeficiente estimado de la parte de latencia, ajustado con una base ortogonal de primer grado del marcador γ: `poly(marker, 1)`.
- Los parámetros de la distribución Weibull:
  - Escala: $\hat{\lambda} = e^{3.56} \approx 35.2$
  - Forma: $\hat{\alpha} = e^{0.586} \approx 1.80$

Se seleccionaron cuatro valores representativos del marcador *ESR1*: 7.5, 9.5, 11.5 y 13.5. Para cada uno de ellos, y para un rango de tiempo de 1 a 175 meses, se calcularon las siguientes funciones:

- $\boldsymbol{h_i(t)}$: función de *riesgo instantáneo*, que mide la probabilidad de recaída en el instante t dado que la paciente no ha recaído hasta entonces.
- $\boldsymbol{S_i(t)}$: función de *supervivencia*, que representa la probabilidad de permanecer libre de recaída hasta el tiempo t.
- $\boldsymbol{f_i(t)}$: función de *densidad*, que indica la probabilidad de recaer exactamente en el tiempo t.

Estas funciones se definieron a partir de los componentes del modelo ajustado:

Tabla 8.4: Definición de las funciones del modelo de cura Weibull con covariable ESR1.

Concepto	**Expresión**	**Interpretación**
Predictor lineal	$\psi_i = \gamma_1 \, z(x_i)$	Modifica el riesgo según el nivel transformado del marcador ESR1.
Riesgo condicional	$h_i(t) = h_0(t) \, e^{\psi_i}$	Probabilidad instantánea de recaída en el tiempo t, entre las pacientes susceptibles.
Supervivencia condicional	$S_i(t) = S_0(t)^{e^{\psi_i}}$	Probabilidad de permanecer libre de recaída hasta el tiempo t para un nivel dado de ESR1.
Densidad condicional	$f_i(t) = h_i(t) S_i(t)$	Probabilidad de que la recaída ocurra exactamente en el instante t.

8.8. Algoritmo para generar representaciones visuales del Modelo de Mezcla de Cura Weibull

Algorithm 3 Generación de las curvas h(t), S(t), f(t)

procedure GENERAR CURVAS(db_205225_at)
  Extraer parámetros del modelo ajustado
  $\gamma_{laten} \leftarrow -7.825$, $\lambda \leftarrow 35.2$, $\alpha \leftarrow 1.80$
  Calcular funciones base Weibull
  $t \leftarrow [1, 2, \ldots, 175]$ ▷ Rango temporal
  $h_0(t) \leftarrow \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}$ ▷ Riesgo base
  $S_0(t) \leftarrow \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right)$ ▷ Supervivencia base
  Seleccionar valores de ESR1
  $markers \leftarrow [7.5, 9.5, 11.5, 13.5]$
  Para cada marcador calcular funciones individuales
  for $m \in markers$ **do**
    $\psi \leftarrow \gamma_{laten} \times poly(m)$ ▷ Efecto del marcador
    $h(t) \leftarrow h_0(t) \times \exp(\psi)$
    $S(t) \leftarrow S_0(t)^{\exp(\psi)}$
    $f(t) \leftarrow h(t) \times S(t)$
  end for
  Generar curvas con ggplot2
  Graficar $h(t)$ vs t para cada marcador
  Graficar $S(t)$ vs t para cada marcador
  Graficar $f(t)$ vs t para cada marcador
end procedure

Para la elaboración de estas curvas se implementó el paquete **ggplot2** y para el manejo de datos de supervivencia se empleó el paquete **survival**, para ajustar el modelo de Kaplan-Meier mediante la función `survfit()` y especificar la estructura de datos censurados con `Surv()`. La transformación polinomial ortogonal del biomarcador ESR1 se realizó con la función `poly()` del paquete base **splines**, garantizando estabilidad numérica en los cálculos. Las funciones de distribución Weibull se calcularon utilizando `dweibull()` para la densidad de probabilidad, `pweibull()` para la función de distribución acumulativa, y combinaciones de estas para derivar las funciones de riesgo y supervivencia base. Finalmente, la optimización de parámetros se llevó a cabo con `nlm()` mediante máxima verosimilitud, y los data frames resultantes se estructuraron para visualizar simultáneamente múltiples curvas diferenciadas por niveles del marcador ESR1.

A continuación, se presentan las curvas generadas para distintos valores del marcador *ESR1*, acompañadas de su interpretación.

8.8.1. Riesgo Instantáneo Condicional

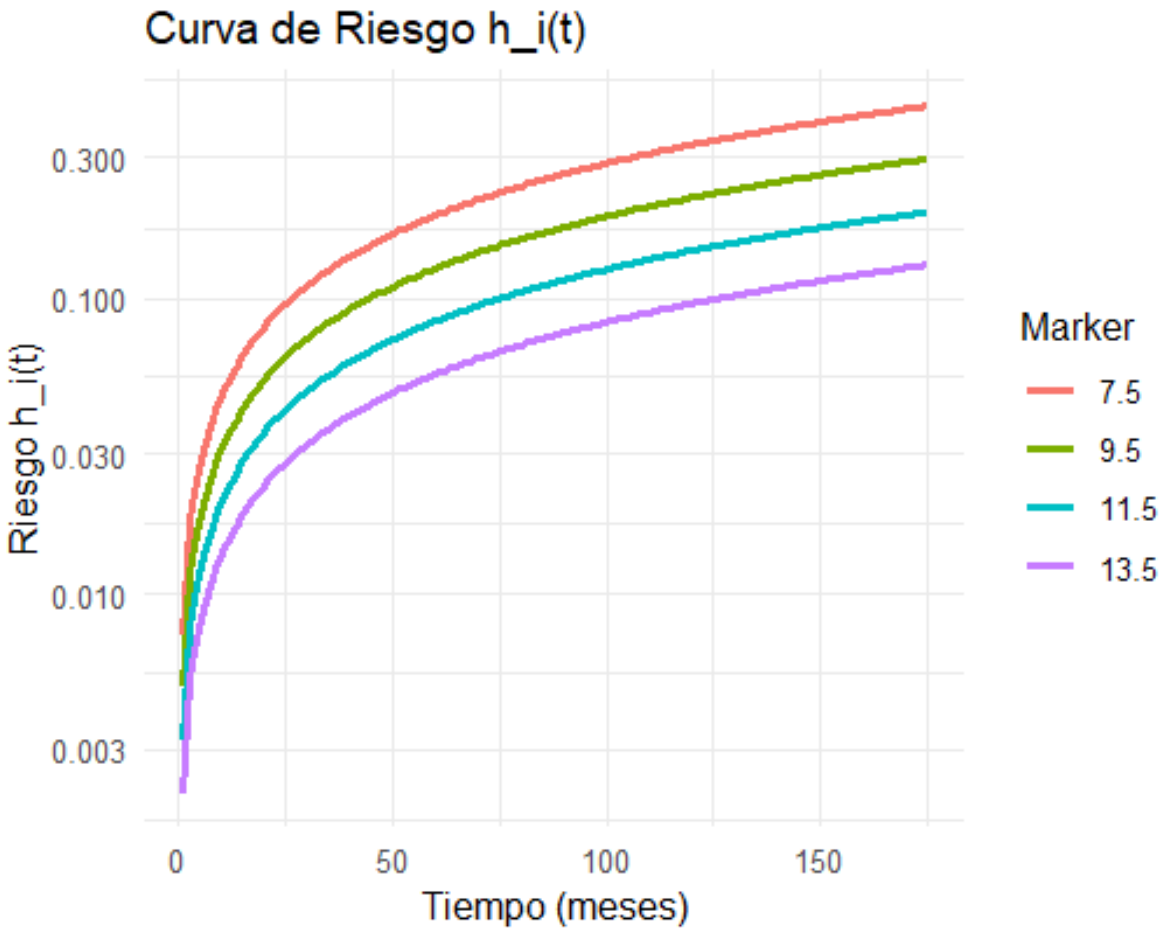


Figura 8.3: Curva de riesgo $h_i(t)$ para diferentes niveles del marcador *ESR1*.

Esta curva muestra la intensidad del riesgo de recaída en cada instante de tiempo, entre las pacientes susceptibles a experimentar el evento se observa que :

- El riesgo de recaída aumenta con el tiempo en todos los grupos, lo cual es consistente con un modelo Weibull con $\alpha > 1$.

- Las mujeres con niveles más bajos de *ESR1* (por ejemplo, 7.5) presentan un mayor riesgo de recaída a lo largo del tiempo, comparadas con quienes tienen niveles más altos (por ejemplo, 13.5).

8.8.2. Supervivencia Condicional

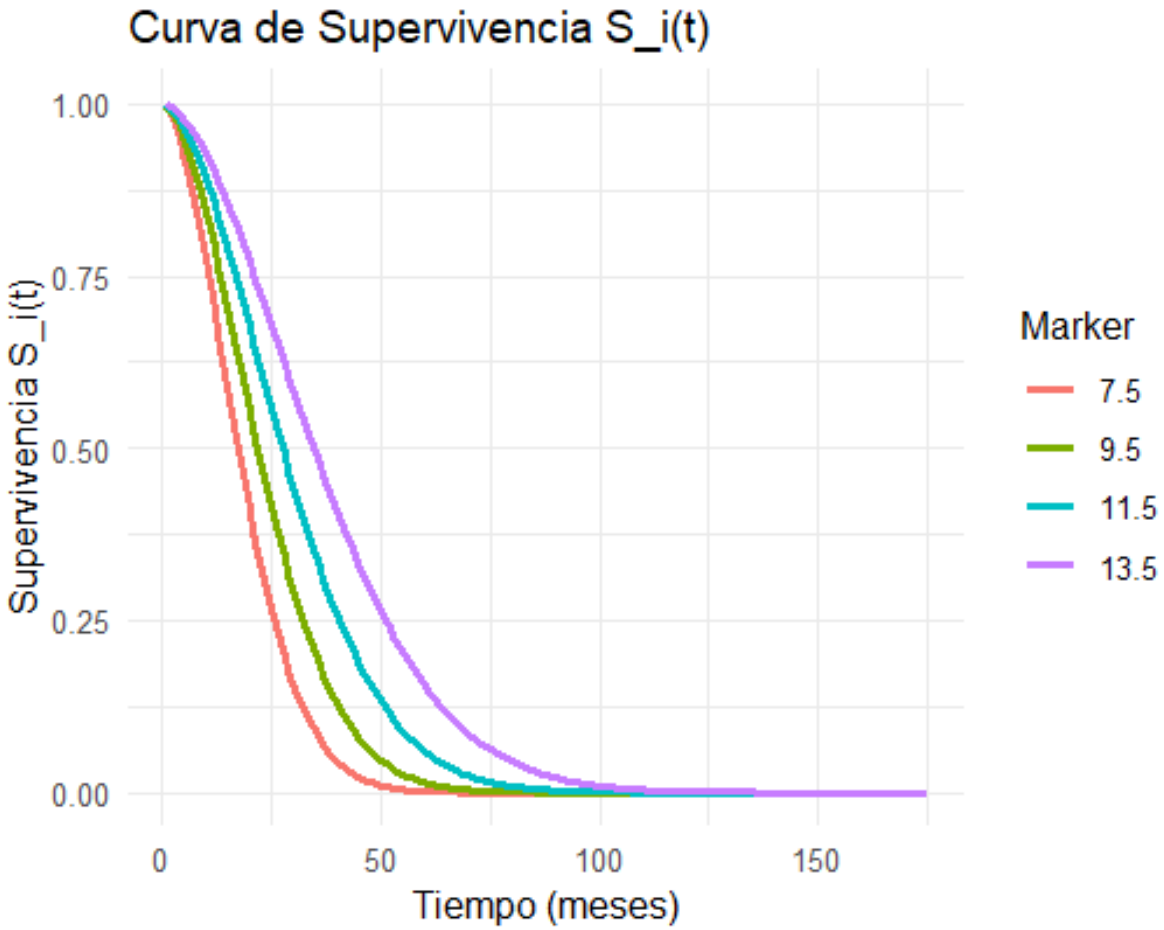


Figura 8.4: Curva de supervivencia condicional $S_i(t)$ para diferentes niveles del marcador *ESR1*.

La curva de supervivencia condicional modela la probabilidad de supervivencia solo dentro del grupo que eventualmente experimentará el evento

Se observa que:

- La probabilidad de mantenerse sin recaer decrece en el tiempo, como se espera en enfermedades progresivas.
- Pacientes con niveles bajos de *ESR1* (por ejemplo, 7.5) recaen rápidamente, con una caída más pronunciada en la curva.
- Pacientes con niveles altos de *ESR1* (por ejemplo, 13.5) mantienen una mayor probabilidad de permanecer libres de recaída durante más tiempo.

Una mayor expresión de *ESR1* se asocia con una mayor duración sin recaída, entre quienes son susceptibles.

8.8.3. Densidad Condicional

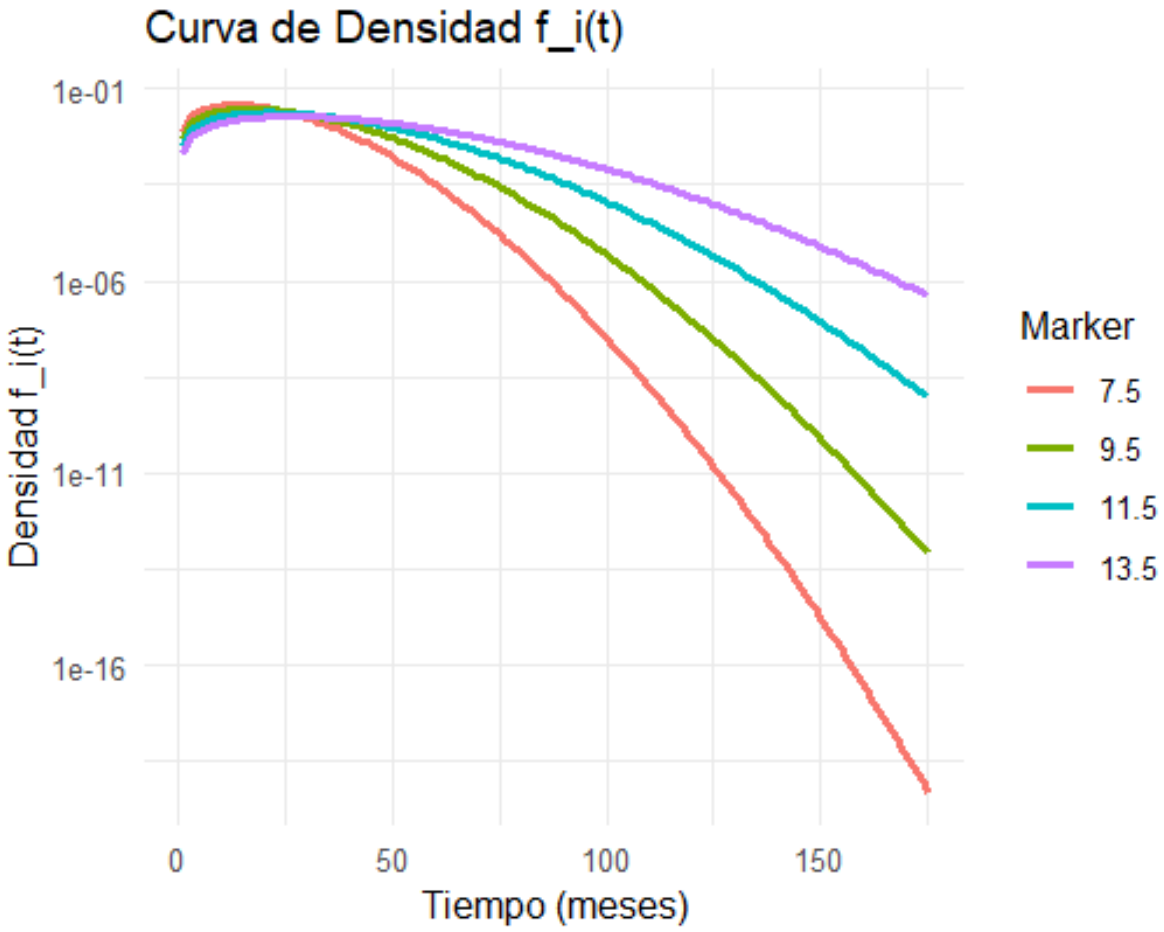


Figura 8.5: Curva de densidad $f_i(t)$ para diferentes niveles del marcador *ESR1*.

La función de densidad refleja el momento más probable en que ocurre la recaída.

Se observa que:

- Para pacientes con niveles bajos del marcador *ESR1* (Por ejemplo, 7.5) la densidad tiene un pico temprano y pronunciado, indicando una alta probabilidad de recaída temprana.
- Para pacientes con niveles altos del marcador *ESR1* el pico es más bajo y desplazado hacia la derecha, indicando una menor probabilidad y recaída desplazada hacia tiempos más tardíos.

Las pacientes con bajo *ESR1* tienen recaídas más rápidas, mientras que niveles altos del marcador están asociados con recaídas más tardías y menos frecuentes.

8.9. Resultados

8.9.1. Comparación del Modelo Nulo vs Modelo Combinado

Tabla 8.5: Comparación entre el modelo nulo y el modelo combinado de cura Weibull.

Parámetro	**Modelo nulo**	**Modelo combinado**	**Interpretación**
Proporción susceptible (logit)	-0.490	-0.240	El modelo combinado ajusta la probabilidad según el nivel del marcador.
Incidencia: marcador < 12.9	—	-0.047 ($p = 0.049$)	Expresión baja de ESR1 reduce la susceptibilidad.
Latencia: efecto del marcador	—	-7.825 ($p < 0.001$)	ESR1 modula fuertemente el tiempo hasta recaída.
$\log(\lambda)$	3.598	3.560	Parámetro de escala de Weibull.
$\log(\alpha)$	0.497	0.586	Parámetro de forma de Weibull.
Log-verosimilitud	Mayor (peor ajuste)	Menor (mejor ajuste)	El modelo combinado explica mejor los datos.

La Tabla 8.5 presenta la comparación entre el modelo nulo y el modelo combinado del modelo de mezcla de cura Weibull. El objetivo es evaluar si la inclusión del marcador *ESR1* mejora la capacidad del modelo para describir la dinámica de recaída en la cohorte estudiada.

En términos generales, se observa que el modelo combinado proporciona un mejor ajuste global, reflejado en una menor log-verosimilitud negativa. Además, los coeficientes asociados al marcador resultan significativos tanto en la parte de incidencia como en la parte de latencia, lo que indica que la expresión génica de *ESR1* influye simultáneamente en la probabilidad de ser susceptible y en el tiempo hasta la recaída entre las pacientes susceptibles.

Estos hallazgos respaldan la pertinencia de incluir *ESR1* en el modelo, ya que aporta información relevante que no está presente en el modelo nulo y permite caracterizar con mayor precisión la heterogeneidad del riesgo entre las pacientes.

8.9.2. Análisis Estratificado por Terciles y Pruebas Log-Rank

Con el objetivo de profundizar en la relación no lineal entre el marcador de expresión génica *ESR1* y el tiempo hasta la recaída, se realizó un análisis de supervivencia estratificado por terciles. El propósito de este enfoque fue validar empíricamente la desición de utilizar un punto de corte, *marker* = 12.9, en el modelo paramétrico y evaluar si el marcador mantiene su capacidad discriminativa dentro de los grupos "Bajoz .Alto"definidos previamente mediante el mismo punto de corte. Para ello, se calculó la distribución de terciles (T1, T2, T3) del marcador de forma separada dentro de cada grupo. Posteriormente, se realizaron comparaciones entre pares de terciles mediante la prueba Log-Rank, evaluando así diferencias en las el riesgo de recaída. Este procedimiento se implementó en `R`, utilizando los paquetes `survival` y `survminer` para estimar y visualizar los modelos.

Resultados para el grupo "Bajo" (marker <12.9)

- Comparación T1 vs T2: $p = 0.261$
- Comparación T1 vs T3: $p = 0.083$
- Comparación T2 vs T3: $p = 0.580$

En el *grupo Bajo*, no se observaron diferencias estadísticamente significativas entre los terciles, aunque la comparación T1 vs T3 sugiere una posible tendencia hacia menor supervivencia en el tercil más bajo T1. En general, se observa un patrón de riesgo relativamente homogéneo, aunque con un cambio en T3.

Resultados para el grupo "Alto" (marker ≥ 12.9*)*

- Comparación T1 vs T2: $p = 0.042$
- Comparación T1 vs T3: $p = 0.142$
- Comparación T2 vs T3: $p = 0.556$

En el *grupo Alto*, se encontró una diferencia significativa entre T1 y T2 ($p = 0.042$), lo que sugiere una mayor capacidad discriminativa del marcador a partir del punto de corte. Esto indica una relación no lineal, donde valores más altos del marcador se asocian con mejor supervivencia. La comparación T2 vs T3 ($p = 0.556$) mostró un cruce de curvas sin significancia, lo que sugiere una posible variación temporal en el riesgo de recaída.

A continuación, se presentan las curvas de supervivencia por terciles del marcador *ESR1* para ambos grupos, junto con los *valores p* obtenidos mediante prueba de log-rank:

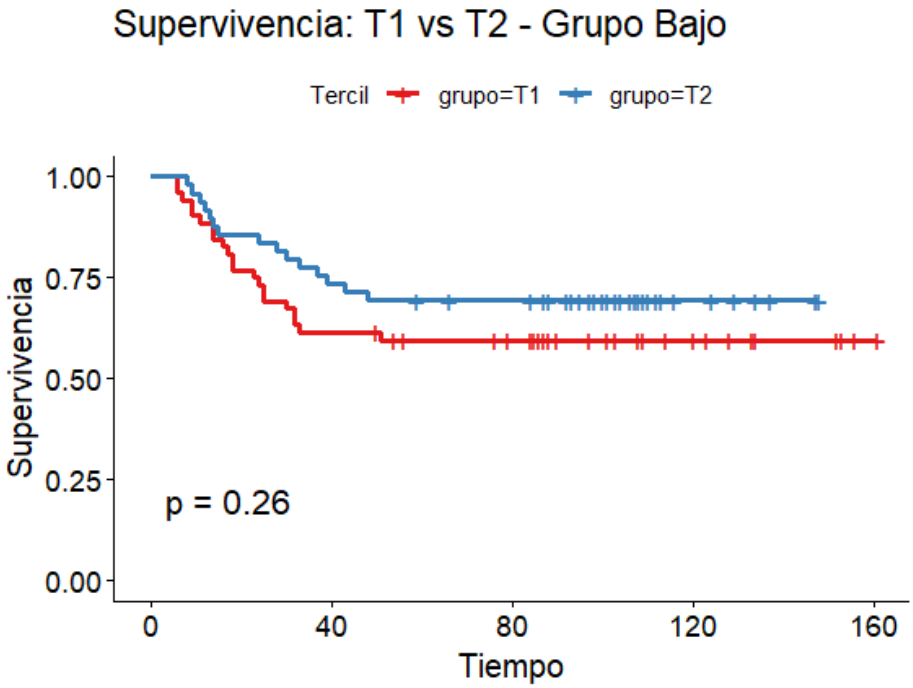


Figura 8.6: Curvas de supervivencia para los terciles T1 (rojo) y T2 (azul) dentro del grupo "Bajo". Aunque se observa una menor supervivencia en T1 en comparación con T2, la diferencia no alcanzó significancia estadística p=0.261.

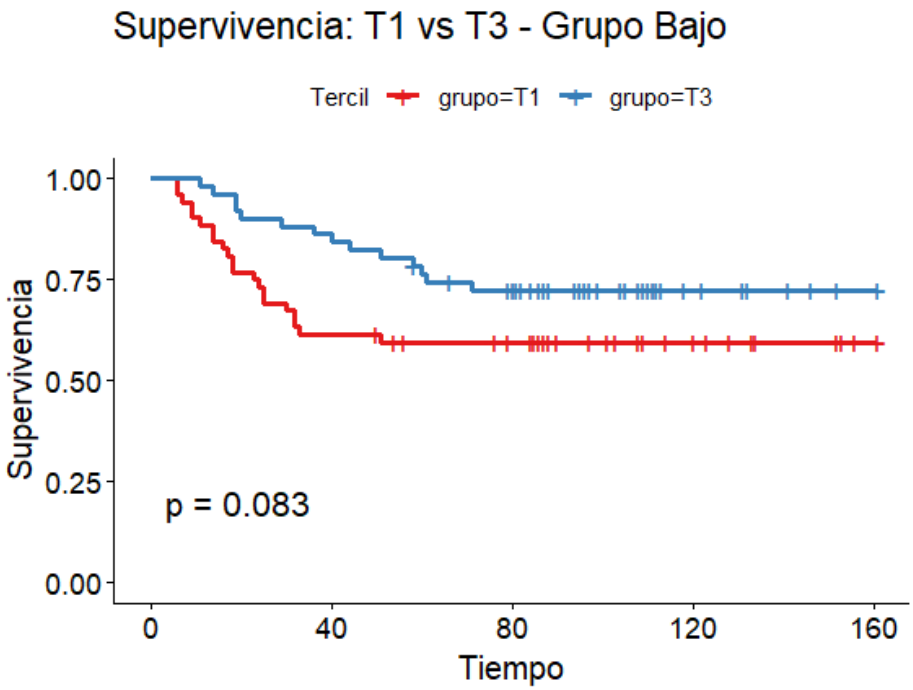


Figura 8.7: Curvas de supervivencia para los terciles T1 (rojo) y T2 (azul) dentro del grupo "Bajo".Se observa una menor supervivencia en T1, con una diferencia no significativa ($p = 0.083$). Sugiriendo que valores más bajos del marcador podrían asociarse con mayor riesgo de recaída, incluso dentro del grupo Bajo.

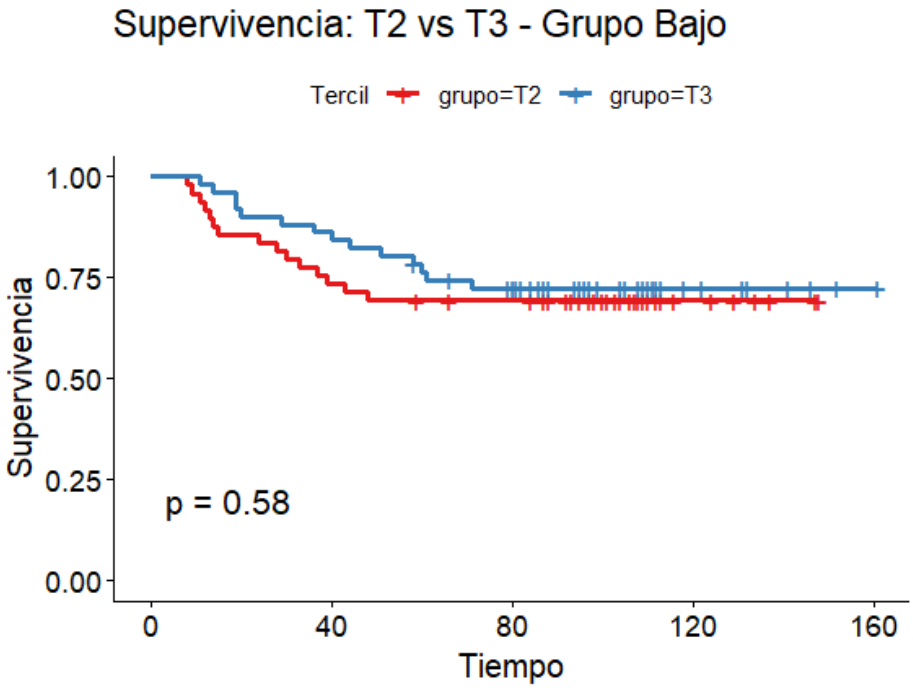


Figura 8.8: Curvas de supervivencia para los terciles T2 (rojo) y T3 (azul) dentro del grupo "Bajo". No se observaron diferencias significativas entre estos subgrupos ($p = 0.580$), lo que sugiere un comportamiento similar en términos de riesgo de recaída.

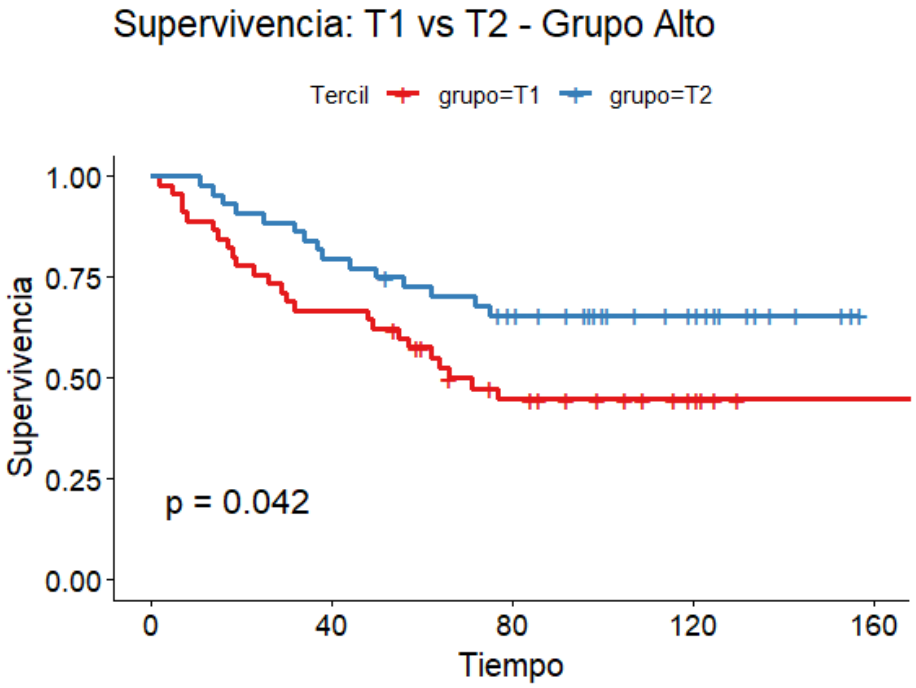


Figura 8.9: Curvas de supervivencia para los terciles T1 (rojo) y T2 (azul) dentro del grupo "Alto". La diferencia entre curvas fue estadísticamente significativa ($p = 0.042$), con mayor supervivencia en T2.

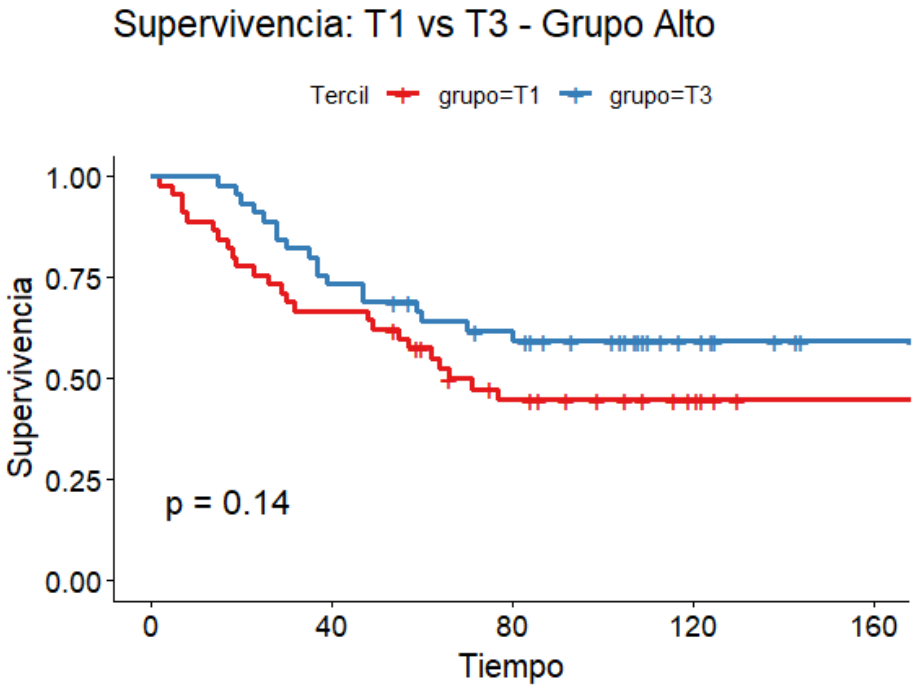


Figura 8.10: Curvas de supervivencia para los terciles T1 (rojo) y T3 (azul) dentro del grupo "Alto". Se observa una tendencia similar a la comparación anterior, aunque sin alcanzar significancia estadística ($p = 0.142$).

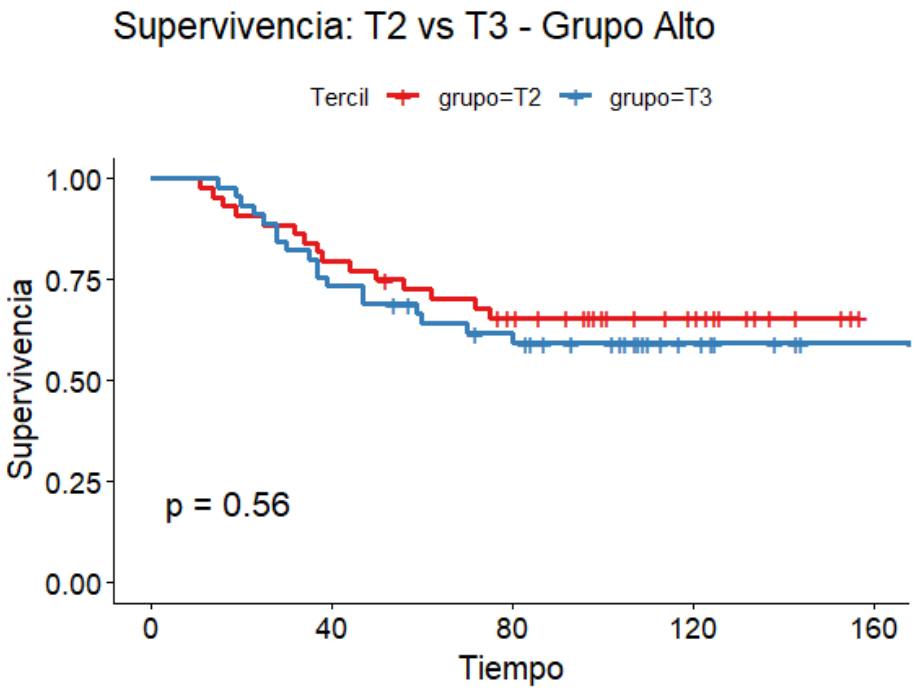


Figura 8.11: Curvas de supervivencia para los terciles T2 (rojo) y T3 (azul) dentro del grupo "Alto". Aunque no se observaron diferencias estadísticamente significativas ($p = 0.556$), las curvas exhiben un cruce: T2 muestra mayor riesgo temprano, mientras que T3 presenta una mayor acumulación de recaídas en el largo plazo.

Estos resultados del análisis por terciles son consistentes con la estructura de función por tramos utilizada en el modelo de mezcla de cura Weibull o también llamado modelo combinado.

La elección de ese umbral se justifica ya que: El grupo bajo presenta un riesgo relativamente homogéneo, aunque con un cambio en la supervivencia y el riesgo cambia abruptamente a partir del punto de corte, donde el marcador adquiere mayor poder discriminativo significativo.
Por tanto, el análisis por terciles no solo permite explorar la relación dentro de cada grupo, sino que además ofrece evidencia complementaria que apoya el modelo estadístico utilizado.

8.9.3. Comparación de la Supervivencia Total estimada vs KM

En la figura 8.12 se muestran las curvas de supervivencia total estimadas por el modelo completo de mezcla de cura Weibull para dos niveles del marcador *ESR1*, definidos según el punto de corte en 12.9. Esta representación permite observar de forma directa las diferencias en la probabilidad de mantenerse libre del evento a lo largo del tiempo, condicionadas al valor del marcador.

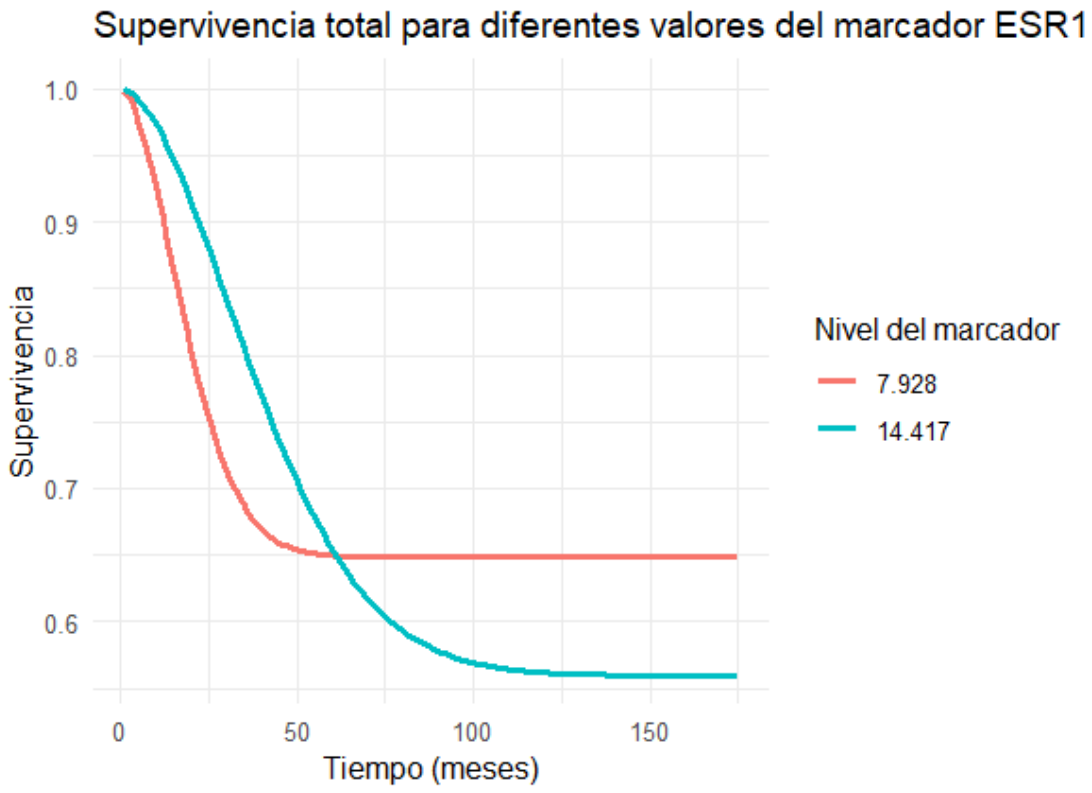


Figura 8.12: Curvas de supervivencia total estimadas por el modelo de cura Weibull para el grupo con nivel bajo (< 12.9) y alto (≥ 12.9) del marcador *ESR1*.

Observamos que la curva correspondiente al valor alto (14.417) desciende más

lentamente, lo que indica una mayor probabilidad de supervivencia. En contraste, para el valor bajo (7.928), la supervivencia disminuye con mayor rapidez. En ambos casos, las curvas no alcanzan el valor cero, lo cual refleja la presencia de una fracción curada estimada por el modelo.

Para validar la capacidad predictiva del modelo, se compararon las curvas estimadas con las curvas empíricas de Kaplan-Meier (KM) correspondientes a los mismos grupos definidos por el punto de corte en 12.9.

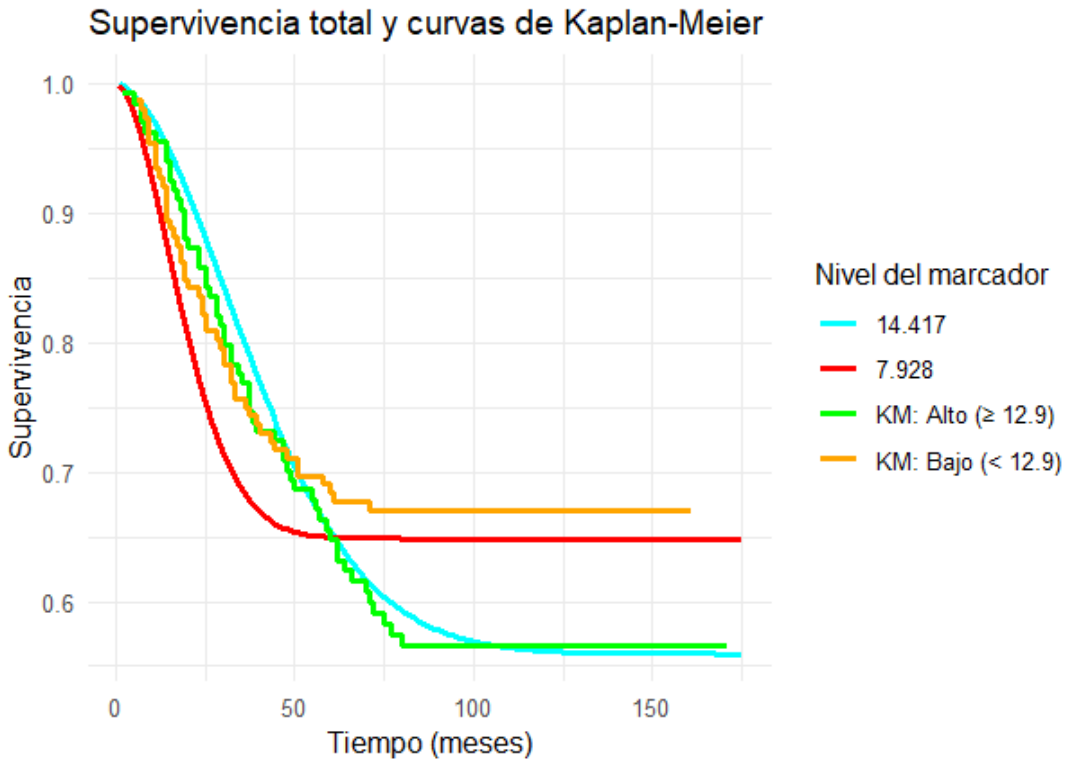


Figura 8.13: Comparación entre las curvas de supervivencia total estimadas por el modelo (rojo y celeste) y las curvas empíricas de Kaplan-Meier (verde y mostaza) para grupos definidos por el punto de corte ESR1 = 12.9.

En la Figura 8.13 las curvas del modelo se ajustan adecuadamente a las tendencias observadas en las curvas de KM, esta concordancia indica que el modelo captura de forma adecuada el patrón de supervivencia observado en los datos. En particular, se mantiene el orden entre grupos, el grupo con niveles altos del marcador presenta una supervivencia prolongada, mientras que el grupo con niveles bajos muestra una caída más acelerada de la probabilidad de supervivencia.

Adicionalmente, se incorporaron los intervalos de confianza al 95 % para las curvas de KM, podrá ver el desarrollo en el anexo, permitiendo comparar la incertidumbre empírica con la predicción puntual del modelo.

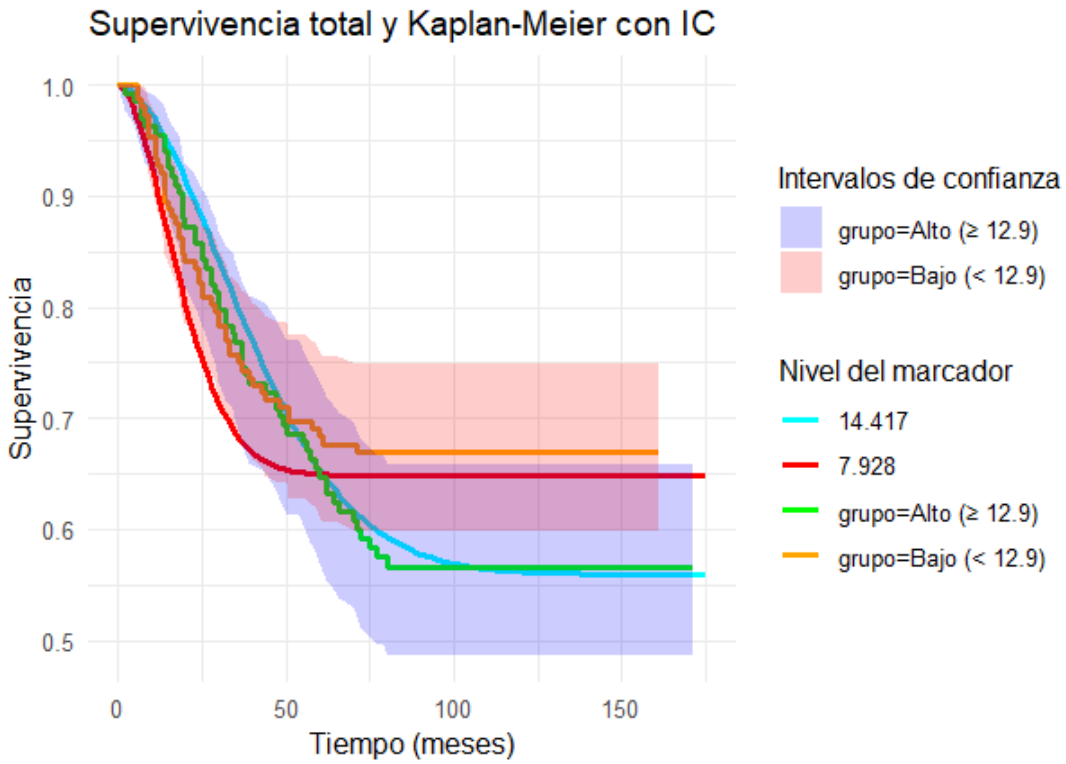


Figura 8.14: Comparación entre las curvas del modelo y las de Kaplan-Meier con intervalos de confianza del 95 % para cada grupo de *ESR1* (bajo: rojo, alto: azul).

En la Figura 8.14 se aprecia que las curvas suaves del modelo de cura Weibull se encuentran dentro o próximas a estos intervalos, lo cual indica una buena concordancia entre el modelo ajustado y los datos observados. Esta consistencia valida el uso del modelo paramétrico y respalda la asociación entre valores elevados del marcador *ESR1* y una mayor probabilidad de supervivencia.

En conjunto, las tres curvas refuerzan la utilidad del marcador *ESR1* como factor pronóstico en pacientes con cáncer de mama, y muestran que el modelo de cura Weibull ajustado logra representar adecuadamente la heterogeneidad en los datos de supervivencia.

Si bien la concordancia visual entre el modelo y las curvas KM es satisfactoria, es importante evaluar de forma cuantitativa el ajuste del modelo. A continuación, se explora a detalle el análisis de los residuos cuantílicos aleatorizados (Randomized Quantile Residuals, RQR) como herramienta diagnóstica, con el fin de validar y detectar posibles desviaciones sistemáticas o problemas de especificación en el modelo ajustado.

8.10. Validación del modelo mediante Residuos Cuantílicos Aleatorizados (RQR)

Una parte fundamental en el desarrollo de modelos estadísticos es la validación de su ajuste. En el contexto de modelos de supervivencia con mezcla de cura, los residuos cuantílicos aleatorizados (RQR, por sus siglas en inglés) constituyen una herramienta poderosa para evaluar la calidad del modelo ajustado. Estos residuos permiten verificar si las predicciones del modelo se ajustan adecuadamente a la distribución teórica, lo que aporta confianza en la validez de las conclusiones obtenidas.

8.10.1. Formulación matemática de los residuos RQR

Los residuos RQR transforman la función de supervivencia estimada $S_i(t_i)$ en una escala normal estándar, permitiendo verificar si el modelo se ajusta bien a los datos bajo la suposición de normalidad de los residuos.

Se definen de la siguiente forma:

$$rqr_i = \begin{cases} \Phi^{-1}\left(\pi_i\left[1 - S_i(t_i)\right]\right), & \text{si el evento fue observado,} \\ \Phi^{-1}\left(1 - \left((1-\pi_i)u_i + \pi_i S_i(t_i)\right)\right), & \text{si fue censurado, } u_i \sim U(0,1), \end{cases}$$

donde:

- $\Phi^{-1}(\cdot)$ es la función cuantil de la normal estándar,
- $S_i(t_i)$ es la supervivencia individual estimada para el tiempo t_i,
- π_i es la probabilidad de ser susceptible estimada para el sujeto i,
- u_i es una variable aleatoria uniforme en (0,1), introducida para aleatorizar el residuo en los casos censurados.

En particular se empleraron las libreías `stats`. Para el cálculo de los residuos cuantil–randomizados (RQR) se emplearon principalmente funciones base de `R`, complementadas con herramientas para el manejo de distribuciones tal como es el caso de la librería `stats`.

Para verificar el supuesto de normalidad se usó la prueba de Anderson-Darling. Un valor de $p = 0.1421$ indicó que no se rechaza la hipótesis nula de normalidad, lo cual respalda la adecuación del modelo.

Algorithm 4 Cálculo de residuos cuantil–randomizados (RQR)

procedure CALCULAR_RQR($t_i,\ status_i,\ \eta_i,\ \psi_i,\ \hat{\lambda},\ \hat{\alpha}$)
Fijar semilla de aleatorización
$n \leftarrow$ número de observaciones

Calcular probabilidad de susceptibilidad
for $i = 1$ **hasta** n **do**
$p_i \leftarrow \text{logit}^{-1}(\eta_i)$
end for

Definir funciones de supervivencia
$S_0(t) \leftarrow 1 - F_{\text{Weibull}}(t \mid \hat{\alpha}, \hat{\lambda})$
$S_i(t) \leftarrow S_0(t)^{\exp(\psi_i)}$

Inicializar RQR
Generar $v_i \sim U(0,1)$
Crear vector vacío rqr

Cálculo de los residuos
for $i = 1$ **hasta** n **do**
Obtener t_i, $status_i$
$S_{i,t} \leftarrow S_i(t_i)$
if $status_i = 1$ **then**
(evento observado)

$$rqr[i] \leftarrow \Phi^{-1}\big(p_i \cdot (1 - S_{i,t})\big)$$

else
(censura con aleatorización)

$$rqr[i] \leftarrow \Phi^{-1}(1 - [(1 - p_i)v_i + p_i S_{i,t}])$$

end if
end for

Almacenar y evaluar
Guardar rqr en la base de datos
Generar histograma y QQ-plot
Aplicar prueba Anderson–Darling
return rqr
end procedure

A continuación, se muestran dos curvas clave para evaluar la distribución de los residuos:

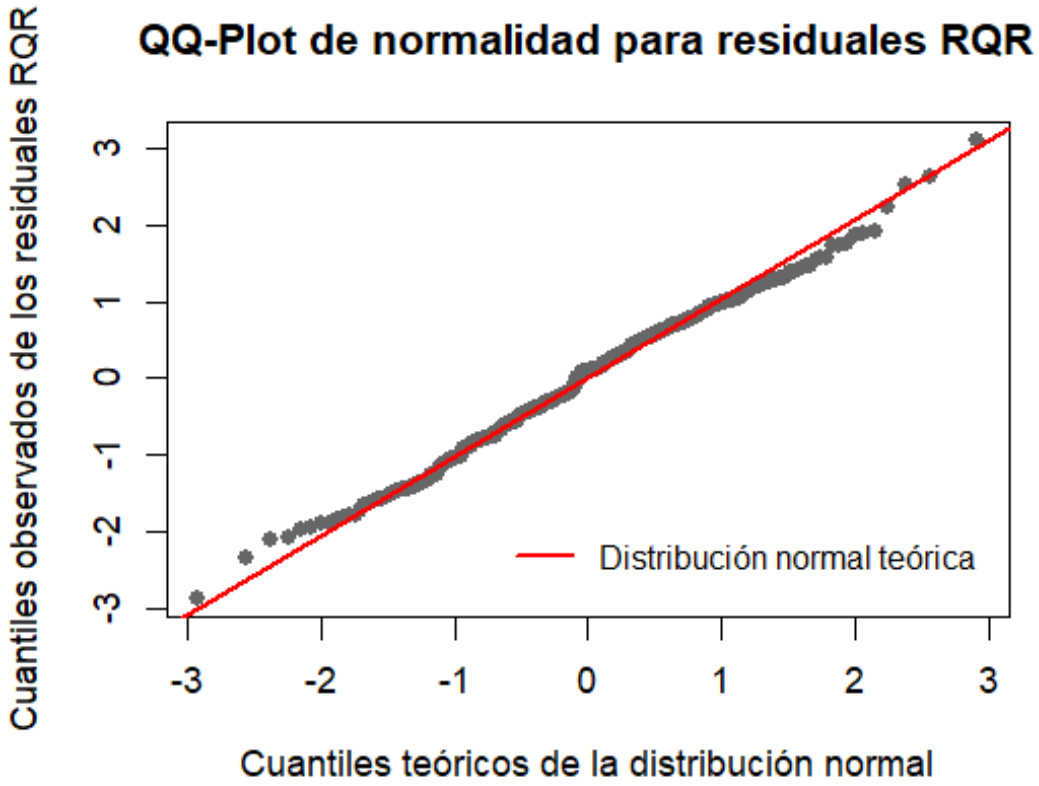


Figura 8.15: Representación Q-Q de los residuos RQR frente a la distribución normal estándar. Se observa que los puntos se alinean razonablemente bien sobre la línea diagonal, lo cual refuerza visualmente el cumplimiento del supuesto de normalidad en los residuos. No se detectan patrones sistemáticos ni colas excesivamente pesadas.

```
Anderson-Darling normality test

data:  db_205225_at$rqr
A = 0.56533, p-value = 0.1421
```

La gráfica QQ-Plot y el test de normalidad de Anderson Darling aplicado a los residuos RQR produjo un estadístico $A = 0.56533$ con un valor $p = 0.1421$ por lo que no se rechaza la hipótesis nula de la normalidad. Indicando así que no existe evidencia estadística para afirmar que los RQR se desvían de la distribución $N(0, 1)$, lo cual es consistente con un ajuste adecuado del modelo.

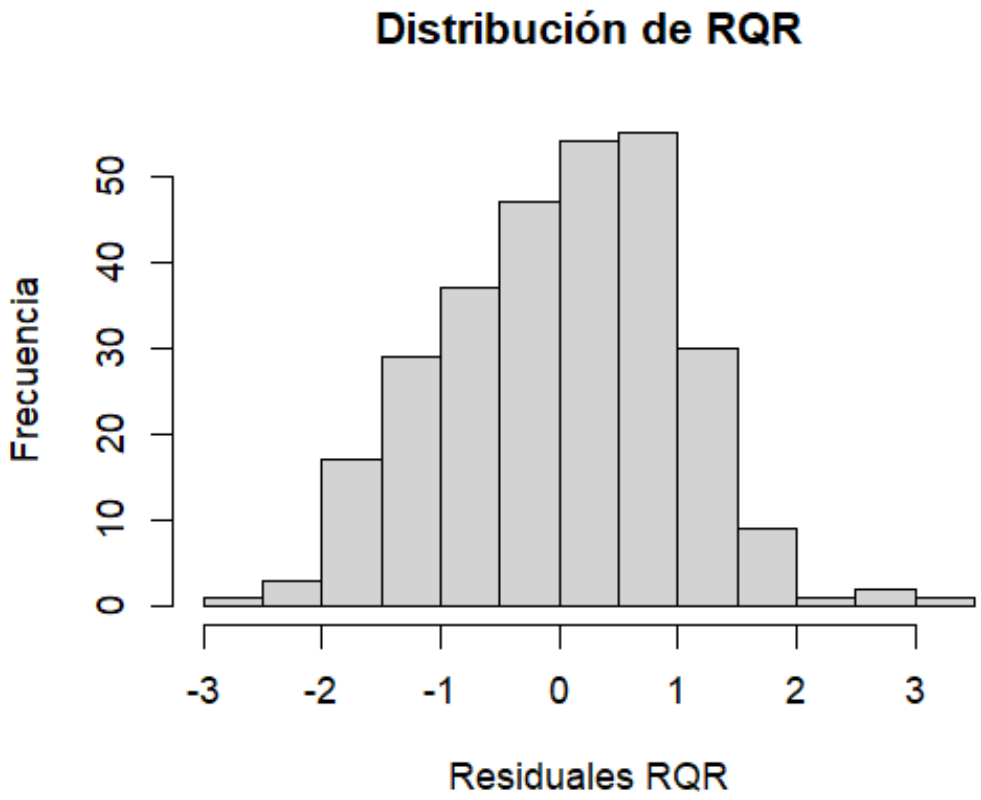


Figura 8.16: Histograma de los residuos RQR estimados a partir del modelo Weibull con marcador *ESR1*, muestra una distribución aproximadamente simétrica y centrada en cero, con forma similar a la normal estándar, lo que sugiere un buen ajuste general del modelo..

El uso de residuos cuantílicos aleatorizados constituye una herramienta robusta para validar modelos de supervivencia con censura, como el modelo de cura Weibull aquí presentado. Los resultados obtenidos confirman su buena especificación y contribuyen a la credibilidad de las estimaciones realizadas. En particular, reafirman el valor del biomarcador *ESR1* para predecir el riesgo de recaída y su utilidad potencial para guiar decisiones clínicas. Además, estos resultados concuerdan con hallazgos previos que sugieren la utilidad de *ESR1* como biomarcador predictivo en cáncer de mama [3].

Con esto concluimos la fase técnica del modelo propuesto. En el siguiente capítulo, se presentan las conclusiones generales de este trabajo, sintetizando los hallazgos principales y discutiendo su relevancia en el contexto clínico y metodológico del análisis de supervivencia.

Capítulo 9

Conclusiones

En este trabajo se logró modelar de manera eficaz la probabilidad de recaída y el tiempo hasta su ocurrencia en pacientes con cáncer de mama en estadio temprano, incorporando la expresión génica del biomarcador *ESR1* como covariable continua. Para ello, se empleó un modelo de supervivencia con fracción curada que combina una función logística para estimar la probabilidad de ser susceptible (recaer) y una distribución Weibull para modelar el tiempo hasta la recaída. Esta estructura permitió capturar de forma adecuada la variabilidad del riesgo a lo largo del tiempo y brindar un enfoque más realista del proceso clínico de recaída, diferenciando tanto la probabilidad de que ocurra el evento como el momento en que se presenta.

Se llevó a cabo la implementación práctica del modelo, su evaluación mediante residuos cuantil-randomizados (RQR) y un análisis detallado del efecto del biomarcador sobre ambas componentes del modelo, apoyado también por visualizaciones de funciones clave, fueron esenciales para respaldar la hipótesis planteada, demostrando que el marcador *ESR1* influye significativamente tanto en la probabilidad de recaída como en el tiempo hasta su ocurrencia. Específicamente, se observó que niveles altos de expresión génica de *ESR1* se asocian con una menor probabilidad de recaída, reflejando una influencia protectora sobre la componente de incidencia del modelo. Esta relación fue claramente capturada por la función logit. Asimismo, *ESR1* mostró un efecto sobre la latencia, modulando la función de riesgo Weibull, lo que confirma su relevancia clínica y estadística en ambas dimensiones del proceso de recaída.

La validación del modelo mediante los residuos RQR aportó evidencia a favor de su adecuación. Los residuos presentaron una distribución aproximadamente normal, y tanto el histograma como el gráfico Q–Q mostraron concordancia con los supuestos del modelo. Esto indica que el modelo Weibull con fracción curada es congruente con la estructura de los datos analizados. Además, el análisis estratificado por terciles del marcador permitió profundizar en la interpretación del efecto del biomarcador en distintos niveles de expresión, reforzando su utilidad como herramienta de estratificación del riesgo.

Las representaciones gráficas de las funciones de supervivencia, riesgo y densidad, derivadas del modelo ajustado y simuladas para distintos niveles del marcador, ofrecieron una interpretación intuitiva y efectiva del comportamiento del proceso de recaída. Así mimo estas herramientas gráficas facilitaron la comunicación de los hallazgos y enriquecieron la interpretación clínica de los resultados.

Desde el punto de vista computacional, se logró una implementación completa del modelo en R, incluyendo funciones específicas para la estimación de parámetros, evaluación del ajuste e interpretación gráfica de los resultados. El uso de transformaciones ortogonales del marcador mediante polinomios, así como transformaciones por tramos contribuyeron a la estabilidad de las estimaciones y demostró ser una técnica eficaz que podría replicarse en contextos con características similares.

En conjunto, los resultados permiten afirmar que se ha cumplido satisfactoriamente el objetivo general del trabajo. El modelo Weibull con fracción de curación no solo fue adecuado para representar la heterogeneidad en los tiempos de recaída, sino que también permitió incorporar de manera flexible el efecto de un marcador molecular continuo, aportando una perspectiva más realista a los estudios de supervivencia. La metodología propuesta se muestra robusta, replicable y útil para futuras investigaciones que involucren datos genómicos y variables clínicas en enfermedades con potencial de curación.

Aunque este estudio se enfocó en el biomarcador *ESR1*, la estrategia metodológica desarrollada puede extenderse a otros genes o firmas moleculares, así como a múltiples covariables clínicas. La posibilidad de adaptar este enfoque a enfermedades con dinámica heterogénea de recaída o curación amplía significativamente su aplicabilidad.

En síntesis, este modelo no solo aporta herramientas estadísticas robustas para la investigación biomédica, sino que también abre el camino hacia una medicina personalizada más precisa, facilitando la identificación de perfiles de riesgo y mejorando la comprensión de procesos complejos en la salud.

Como línea futura, podría considerarse la incorporación de modelos más flexibles, tales como los enfoques semiparamétricos o bayesianos, que permitirían captar estructuras no lineales y distribuciones más generales en escenarios clínicos diversos. Sin embargo, el enfoque desarrollado en este trabajo ya constituye una base sólida para la exploración del comportamiento de biomarcadores en estudios de supervivencia con fracción de curación.

Bibliografía

[1] Amico, M., & Van Keilegom, I. (2018). *Cure Models in Survival Analysis: From Modelling to Prediction Assessment of the Cure Fraction*. Tesis doctoral, Catholic University of Louvain (UCL) y KU Leuven.

[2] Brett, J. O., Spring, L. M., Bardia, A., Wander, S. A., & Rugo, H. S. (2021). ESR1 mutation as an emerging clinical biomarker in metastatic hormone receptor-positive breast cancer. *Breast Cancer Research*, *23*(1), 85. https://doi.org/10.1186/s13058-021-01462-3

[3] Budczies, J., Klauschen, F., Sinn, B. V., Győrffy, B., Schmitt, W. D., Darb-Esfahani, S., & Denkert, C. (2012). Cutoff Finder: A comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PLOS ONE*, *7*(12), e51862. https://doi.org/10.1371/journal.pone.0051862

[4] Cai, C., Zou, Y., Peng, Y., & Zhang, J. (2013). smcure: An R-package for estimating semiparametric mixture cure models. *Journal of Statistical Software*, *54*(11), 1–23. https://pmc.ncbi.nlm.nih.gov/articles/PMC3494798/

[5] Collett, D. (2003). *Modelling survival data in medical research* (2nd ed.). Chapman & Hall/CRC.

[6] Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. Chapman & Hall/CRC.

[7] Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244. https://doi.org/10.2307/1390802

[8] Escarela Pérez, G. (2006). *Modelos de supervivencia*. Universidad Autónoma Metropolitana, Departamento de Matemáticas. (Docencia 04.0405.11.001.2006).

[9] Herring, A. H., & Ibrahim, J. G. (2002). Maximum likelihood estimation in random effects cure rate models with nonignorable missing covariates. *Biostatistics*, *3*(3), 387–405. https://doi.org/10.1093/biostatistics/3.3.387

[10] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.
https://doi.org/10.1080/01621459.1958.10501452

[11] Maller, R. A., & Zhou, X. (1996). *Survival analysis with long-term survivors.* Wiley.

[12] National Cancer Institute. (s.f.). *Affymetrix U133 data sheet.* National Institutes of Health.
https://www.cancer.gov/ccg/research/structural-genomics/tcga/using-tcga-data/technology/affymetrix-u133-data-sheet

[13] National Center for Biotechnology Information. (2005). Gene expression profile data from breast cancer patients (GSE2034) [Data set]. *Gene Expression Omnibus.*
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034

[14] Organización Panamericana de la Salud. (2023). *Cáncer de mama en las Américas.*
https://www.paho.org/es/temas/cancer-mama

[15] Othus, M., Barlogie, B., LeBlanc, M. L., & Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, *18*(14), 3731–3736.
https://doi.org/10.1158/1078-0432.CCR-11-2859

[16] Peng, Y., & Yu, B. (2021). *Cure Models: Methods, Applications, and Implementation* (1st ed.). Chapman and Hall/CRC. New York. https://doi.org/10.1201/9780429032301

[17] Terán-Hernández, M., González-Arias, M., Garrocho-Rangel, C. F., & Ruiz-Rodríguez, M. d. l. Á. (2020). A survival analysis with time-dependent covariates: A versatile simulation tool based on the permutation of failure times. *BMC Medical Research Methodology*, *20*(1), Article 163.
https://doi.org/10.1186/s12874-020-01055-2

[18] Vargas-García, C. F. (2021). *Supervivencia y series: Material del curso FC2021-1* [Notas de curso]. GitHub Pages.
https://carlosfernandovg.github.io/supervivencia_y_series_FC2021-1/index.html

[19] Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., & Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, *365*(9460), 671–679.
https://doi.org/10.1016/S0140-6736(05)17947-1

Anexos

A.0.1. Estimación de intervalos de confianza en la función de incidencia mediante simulación Monte Carlo

Para evaluar la incertidumbre en la estimación de la probabilidad de recaída (susceptibilidad), se recurrió a una aproximación Monte Carlo basada en la matriz de varianza-covarianza del estimador de máxima verosimilitud. La formulación matemática es la siguiente:

- Sean $\hat{\boldsymbol{\beta}}$ los coeficientes estimados del modelo de incidencia, y $\hat{\Sigma}$ la matriz de varianza-covarianza asociada.
- Se simulan $M = 1000$ vectores $\boldsymbol{\beta}^{(m)} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$, para $m = 1, \ldots, M$.
- Para una secuencia de valores del marcador x, se calcula:

$$\eta^{(m)}(x) = \mathbf{x}^\top \boldsymbol{\beta}^{(m)}, \quad \text{y} \quad \pi^{(m)}(x) = \frac{1}{1 + e^{-\eta^{(m)}(x)}}$$

- Los intervalos de confianza se obtienen como los percentiles 2.5 y 97.5 de la distribución empírica de $\pi^{(m)}(x)$.

El siguiente fragmento en R implementa esta metodología:

```
# Simulación de Monte Carlo para intervalos de confianza del logit(p)
set.seed(123)
n_sim <- 1000
Z <- matrix(rnorm(n_sim * length(beta_incid)), nrow = n_sim)
L <- chol(Sigma_incid)  # Descomposición de Cholesky de la var-cov
beta_sim <- sweep(Z %*% L, 2, beta_incid, FUN = "+")

# Evaluamos eta para cada simulación y valor del marcador
eta_sim <- new_mat_incid %*% t(beta_sim)

# Calculamos los percentiles 2.5% y 97.5% para eta
eta_lower <- apply(eta_sim, 1, quantile, probs = 0.025)
eta_upper <- apply(eta_sim, 1, quantile, probs = 0.975)
eta_med   <- apply(eta_sim, 1, median)

# Transformación inversa logit para obtener pi
p_lower <- plogis(eta_lower)
p_upper <- plogis(eta_upper)
p_med   <- plogis(eta_med)
```

Estos resultados se utilizaron para graficar los intervalos de confianza en la curva de logit (η) y la curva de probabilidad de ser susceptible (π) en función del marcador. Este método permite capturar la forma y la variabilidad del modelo sin depender de aproximaciones lineales.

B.0.2. Estimación de intervalos de confianza para el estimador de Kaplan-Meier

Para las curvas empíricas de supervivencia estimadas mediante el método de Kaplan-Meier, se calcularon intervalos de confianza utilizando la transformación log–log y la varianza de Greenwood, cuya formulación matemática es:

$$\widehat{\mathrm{Var}}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

donde:

- d_i: número de eventos en el tiempo t_i
- n_i: número de individuos en riesgo justo antes de t_i
- $\hat{S}(t)$ es la estimación de la función de supervivencia

Los intervalos de confianza al 95 % se obtienen mediante la transformación log-log:

$$\left[\hat{S}(t)^{\exp(z_{\alpha/2} \cdot \frac{\sqrt{\widehat{\mathrm{Var}}[\hat{S}(t)]}}{\hat{S}(t) \log \hat{S}(t)})},\ \hat{S}(t)^{\exp(-z_{\alpha/2} \cdot \frac{\sqrt{\widehat{\mathrm{Var}}[\hat{S}(t)]}}{\hat{S}(t) \log \hat{S}(t)})} \right]$$

En R, se utilizó la función `survfit` de `survival`, que por defecto utiliza la varianza de Greenwood y la transformación log–log y `ggsurvplot` de `survminer` para obtener las curvas y sus bandas de confianza:

```
# Ajuste Kaplan-Meier
km_fit <- survfit(Surv(time, status) ~ grupo, data = db_205225_at)

# Graficar KM con IC
gg_km <- ggsurvplot(km_fit, conf.int = TRUE)
km_data_full <- gg_km$plot$data

ggplot(df_plot, aes(x = t, y = S_total, color = as.factor(marker)))+
  geom_line(size = 1.2) +
  # Intervalos de confianza
  geom_ribbon(data = km_data_full,
              aes(x = time, ymin = lower, ymax = upper,
              fill = strata),
              alpha = 0.2, inherit.aes = FALSE) +
```

Con ello, se visualiza la función de supervivencia estimada junto con los intervalos de confianza que reflejan la incertidumbre en la estimación en cada punto del tiempo.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00251
Matrícula: 2233803393

Modelado de datos de supervivencia con una fracción no susceptible.

En la Ciudad de México, se presentaron a las 10:30 horas del día 15 del mes de diciembre del año 2025 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. GABRIEL NUÑEZ ANTONIO
DRA. ELIZABETH GONZALEZ ESTRADA
DR. GABRIEL ESCARELA PEREZ

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRA EN CIENCIAS (MATEMÁTICAS APLICADAS E INDUSTRIALES)

DE: CLAUDIA CHAPARRO VELAZQUEZ

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, el presidente del jurado comunicó a la interesada el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

CLAUDIA CHAPARRO VELAZQUEZ
ALUMNA

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. ROMAN LINARES ROMERO

PRESIDENTE

DR. GABRIEL NUÑEZ ANTONIO

VOCAL

DRA. ELIZABETH GONZALEZ ESTRADA

SECRETARIO

DR. GABRIEL ESCARELA PEREZ