UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

TRABAJO FINAL DE GRADO PARA OBTENER EL TÍTULO DE LICENCIADO EN ESTADÍSTICA

VIAR: Una herramienta para la estimación de la función de riesgo.

Diego Araújo Arellano

Tutor:

Dr. Marco Scavino

Montevideo

URUGUAY

2022

FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

El tribunal docente integrado por los abajo firmantes aprueba el Trabajo Final de Grado:

VIAR - Una herramienta para la estimación de la función de riesgo.

Diego Araújo Arellano

Tutor:

Dr. Marco Scavino

Licenciatura en Estadística

Puntaje 12

Tribunal

Profesor Dr. Ramón Álvarez-Vaz.....(firma).

Profesor Dr. Daniel Ciganda.....(firma).

Profesor Dra. Natalia da Silva.....(firma).

Profesor Dr. Marco Scavino.....(firma).

Fecha 24/06/2022

Resumen

El estudio de métodos para analizar datos sobre eventos observados a lo largo del tiempo y de factores asociados con la tasa de ocurrencia de estos eventos, es un tema que siempre ha preocupado, sobre todo en el ámbito de la salud. De esta forma, el análisis se le conoce genéricamente como "Análisis de supervivencia". Si bien por lo general se hace énfasis en la función de supervivencia en este tipo de análisis, poder estimar de la mejor manera la tasa de riesgo instantánea asociada a alguna enfermedad hasta el fallo, remisión o recurrencia de la misma, es también un tema relevante.

Si bien se pueden utilizar varios modelos paramétricos en este tipo de estimaciones (Exponencial, Weibull y Gamma entre otros), el riesgo de utilizar estos modelos está en cometer un error de especificación del modelo, lo que podría derivar en resultados incorrectos. Para esto podemos hacer uso de modelos no paramétricos, como es la estimación mediante núcleos (*kernel*), donde no realizamos ninguna suposición inicial sobre la distribución de los datos, sino que la única información es proporcionada por la muestra.

El eje principal de este trabajo es crear una herramienta, para estimar la función de riesgo mediante núcleos, a través de una aplicación web (`shiny`). Para ello, se utilizó como base el trabajo realizado por los autores de la Universidad de Mazaryk (Selingerová, Doleželová et al. 2016).

"VIAR" - como se le ha llamado a la herramienta - pretende ser un aplicación interactiva, con la cuál se puede realizar el análisis de supervivencia (y en particular de la función de riesgo) de una manera bastante simple, evitando al usuario tener que lidiar con el código necesario para hacerlo. Se puede acceder a la misma a través del siguiente link: https://diegoarare.shinyapps.io/VIAR.

A lo largo del informe se explicarán las metodologías utilizadas, para luego pasar al análisis de una base de datos determinada, siempre teniendo en cuenta que el trabajo está enfocado a la creación de un producto y fue pensado en cada instancia como tal. Se hará uso del lenguaje de programación R y el entorno de desarrollo Rstudio, para la creación de la herramienta.

Palabras Clave: Análisis de supervivencia, estimación mediante núcleos (kernel), función de riesgo univariada, función de riesgo condicional, modelo de Cox.

Agradecimientos

En primera instancia quisiera agradecer a mi tutor, Marco Scavino, quien con dedicación me acompañó en esta etapa final de la carrera, brindando su apoyo y conocimiento para lograr este trabajo final de grado. A Ramón Álvarez-Vaz, que no tuve la suerte de tenerlo como docente durante la carrera, pero si de trabajar en algún proyecto de investigación y continuamente me alienta a superarme. A Silvia Rodríguez, quien me dió ese "empujoncito" para acercarme a la investigación. A todos los profesores y compañeros (y compañeros-profesores) con quienes compartí y contribuyeron de alguna manera a mi formación a lo largo de la carrera.

A Iveta Selingerová, que se tomó el tiempo para contestar mi consulta y logró sacarme el bloqueo que tenía en una etapa de la programación.

A mi familia que fue un apoyo importante, sobre todo en los últimos años que fueron un poco más difíciles y siempre estuvieron ahí, apoyándome, escuchándome, dándome aliento. A Ceci, que me viene acompañando todo este camino, bancándome siempre. A mis amigos, que siguen sin entender que hago ("esas fórmulas raras...") pero siempre me están apoyando, festejando las buenas y levantándome en las malas.

Por último quiero agradecer a la persona que me dió fuerza para seguir siempre (y que de alguna manera lo sigue haciendo). Mi hermana, *Vicky*, quien me acompañó desde que tengo uso de razón y siempre me impulsó a ser mejor. Gracias por acompañarme, cuidarme, enseñarme y por todo lo que me dejaste. Me gustaría que estuvieras acá para compartir este momento.

La aplicación lleva las siglas de su nombre: Virginia Araújo.

Índice

Índice de figuras

1. Introducción

El estudio de métodos para analizar datos sobre eventos observados a lo largo de un tiempo y de los factores asociados con la tasa de ocurrencia de estos eventos, es un tema que ha tomado considerable atención a lo largo de los años.

Son diversos los campos que se nutren de estos métodos para analizar dichos conjuntos de datos, como pueden ser la Medicina, la Biología, la Epidemiología, Ingeniería, Demografía, Ciencias Actuariales y la Economía. Los datos estudiados se llaman de forma genérica *"datos de supervivencia"*, debido a que este tipo de análisis en sus inicios era utilizado para observar la evolución de pacientes con determinada enfermedad hasta la ocurrencia del fallecimiento.

El objetivo del análisis de supervivencia es conocer, analizar, y predecir los tiempos de duración en una determinada situación que termina con la ocurrencia de un evento. Para esto se estudia el tiempo que transcurre desde un estado inicial, por ejemplo el diagnóstico de cierta enfermedad, hasta que sucede el evento de interés, como puede ser el fallecimiento del individuo.

Ejemplos donde se aplican este tipo de análisis son:

- En estudios referidos a tumores:
  - Tiempo desde la cirugía de un tumor hasta el fallecimiento
  - Tiempo de inicio del tratamiento hasta remisión
  - Tiempo de recurrencia de la enfermedad
- Tiempo de infección de VIH hasta desarrollo del Sida
- Tiempo de infarto
- Tiempo de inicio de abuso de sustancias
- Tiempo de mal funcionamiento de una máquina (o rotura de una pieza)

Hay tres funciones que caracterizan la distribución del tiempo de duración hasta cierto evento. Por un lado tenemos la *función de supervivencia*, que es la probabilidad de que un individuo sobreviva a un tiempo x. Luego tenemos la *función de riesgo acumulado*, que tiene una relación inversa con el logaritmo de la supervivencia. Por último tenemos la *función de riesgo*, que representa la tasa de mortalidad instantánea (o de fallo, según el contexto) para un individuo que sobrevive al tiempo x. Esta última ha recibido relativamente menos atención y es en la cuál se enfocará el trabajo.

Los modelos paramétricos son ampliamente utilizados para estimar estas funciones, no solo por la cantidad de investigaciones existentes al respecto, sino también porque ofrecen información sobre la naturaleza de los diversos parámetros. Algunos de estos modelos incluyen por ejemplo, la distribución Exponencial, Weibull, Gamma, Lognormal, Pareto, etc. El riesgo de utilizar estos modelos está en cometer un error de especificación del modelo, que podría derivar en resultados incorrectos.

Para esto nos enfocamos en modelos no paramétricos, donde no realizamos ninguna suposición inicial sobre la distribución de los datos, sino que la única información es proporcionada por la muestra.

El trabajo se estructura en 7 secciones. En la sección 2 se presenta el objetivo principal del trabajo. Siguiente a esto, en la sección 3 se establece un marco teórico con definiciones comúnmente utilizadas en el análisis de supervivencia y que se manejarán a lo largo del documento. La sección 4 introduce en la metodología estadística utilizada por la herramienta. La sección 5 presenta un detalle de los paquetes y librerías utilizadas para la programación de la aplicación y un manual de uso que indica los pasos a seguir para realizar el análisis, el cúal se realizará en la sección 6, tomando como prueba un conjunto de datos de supervivencia. Por último en la sección 7 se presentan las conclusiones, comentarios finales y el posible trabajo futuro que genera la creación de esta herramienta.

2. Objetivo del trabajo

El objetivo principal del trabajo es crear una herramienta de manejo simple, para el análisis de supervivencia y en particular de la función de riesgo. La misma consistirá en una aplicación web usando el paquete `shiny` (Chang, Cheng et al. 2021), creada a través del lenguaje `R` (R Core Team 2020).

Se abordará como metodología, la estimación de la función de riesgo a través de métodos de kernel establecidos por (Horová, Kolácek y Zelinka 2012) y luego abordado además por (Selingerová, Doleželová et al. 2016). En primera instancia se adaptará el riesgo univariado, es decir del estudio del tiempo hasta la ocurrencia del evento de interés, para luego extender la estimación de la función de riesgo al uso de alguna covariable de interés.

También se creará una sección para modelos de uso frecuente, como lo es el sugerido por Cox (Cox 1972) (modelo semiparamétrico), para luego comparar ambos métodos con sus virtudes y limitaciones.

3. Marco Teórico (MT)

Antes de introducirnos en la metodología, comenzaremos con algunas definiciones que se manejarán en el documento y otras propias del análisis de supervivencia.

3.1. Tiempo de vida

El tiempo de vida o de supervivencia, representado por la variable aleatoria T (continua y no negativa), se define como el tiempo transcurrido hasta que se produce algún evento de interés. Generalmente nos referimos a este como el tiempo desde el comienzo del seguimiento (del estudio) hasta la muerte o falla del mecanismo.

3.2. Función de supervivencia

Llamamos F a la función de distribución de T, es decir $F(x) = P(T < x)$. Por lo que el proceso supervivencia puede caracterizarse por la función de supervivencia $\overline{F}$:

$$\overline{F}(x) = P(T \geq x) = 1 - F(x) \tag{1}$$

La misma se define como la probabilidad de que el tiempo de supervivencia sea mayor o igual a x.

3.3. Función de riesgo

La función de riesgo es la probabilidad de que un individuo muera (o falle) en el tiempo x, condicionado a que haya sobrevivido hasta ese momento. Si la distribución de vida F tiene densidad f, para $\overline{F}(x) > 0$, la función de riesgo es definida por

$$\lambda(x) = \frac{f(x)}{\overline{F}(x)} \tag{2}$$

y la función de riesgo acumulado como

$$H(x) = -log\overline{F}(x) \tag{3}$$

La función de riesgo es una herramienta útil en el análisis de supervivencia ya que refleja la probabilidad instantánea de que se produzca el fallo. En la práctica, si bien nos interesa analizar tiempos hasta el evento, por lo general esta función depende de covariables, como puede ser la edad o el sexo si estamos aplicando en áreas de la Medicina, o por ejemplo del tipo de material si hablamos de piezas industriales.

La función de supervivencia y la de riesgo, son complementarias, nos ofrecen distintos puntos de vista de los datos analizados.

3.4. Tipos de Censura

Otro elemento a tomar en consideración al momento de realizar un análisis de supervivencia, es que nos encontramos con diferentes fenómenos que afectan a los datos como pueden ser la censura y el truncamiento, los cuales se explicarán en los siguientes puntos.

La censura es un fenómeno que ocurre cuando el valor de una observación sólo se conoce de forma parcial. Existen diferentes tipos de censura, los cuales se detallan a continuación.

3.4.1. Censura por la derecha

Se presenta cuando una vez finalizado el periodo de observación de un determinado individuo, éste aún no ha presentado el evento que pretendemos observar. Por ejemplo, se observa en estudios donde el investigador debe determinar un tiempo fijo de observación de los n sujetos Cr. En estas circunstancias, puede ocurrir que a la finalización del estudio, todos los sujetos de la muestra no hayan presentado el suceso final.
En este caso, los datos registrados se pueden representar usando un par de variables aleatorias (Y, δ), donde:

$$\delta = \begin{cases} 1 & \text{si el evento se ha observado} \\ 0 & \text{si el evento es censurado,} \end{cases}$$

$$Y = \begin{cases} T & \text{si el tiempo de vida es observado} \\ C_r & \text{si el tiempo de vida es censurado,} \end{cases}$$

o lo que es lo mismo $Y = min\,(T, C_r)$.

3.4.2. Censura por la izquierda

Un tiempo de vida T, asociado a un individuo específico se considera censurado por la izquierda cuando el evento de interés ha ocurrido antes de que este sea incluido en el estudio.
Para estos individuos, conocemos que han experimentado el evento en algún momento anterior a C_l, pero desconocemos el momento exacto en que se ha producido, ya que esto solo sería conocido si $T \geq C_l$.
Estos datos se pueden representar por el par de variables aleatorias (Y, ϵ), donde:

$$\epsilon = \begin{cases} 1 & \text{si el evento se ha observado} \\ 0 & \text{si el evento es censurado,} \end{cases}$$

$$Y = \begin{cases} T & \text{si el tiempo de vida es observado} \\ C_l & \text{si el tiempo de vida es censurado,} \end{cases}$$

o lo que es lo mismo $Y = max\,(T, C_l)$.

3.4.3. Censura doble

Algunos estudios presentan datos censurados por la izquierda y por la derecha, con lo que sus tiempos de vida se consideran que están doblemente censurados. Aquí, también se pueden representar los datos por una pareja de variables (Y_i, δ_i), donde:

$$Y = max\,[min\,(T, C_r)\,, C_l]$$

$$\delta_i = \begin{cases} 1 & \text{si Y es un evento observado} \\ 0 & \text{si Y es un dato censurado por derecha} \\ -1 & \text{si Y es un dato censurado por izquierda,} \end{cases}$$

Por lo que C_l es un tiempo anterior al comienzo del estudio en el que los individuos experimentan el evento, C_r es un tiempo posterior al momento de finalización del estudio y solo se conocerá si $T \leq C_r$ y $T \geq C_l$.

3.4.4. Censura aleatoria

Este tipo de censura se presenta cuando en el transcurso de un estudio, las observaciones experimentan otro tipo de suceso independiente del evento de interés que provoca su salida del mismo. En estudios de supervivencia esto podría darse, por ejemplo, por muertes con causas ajenas al motivo de estudio, por pacientes que abandonan la prueba clínica por un cambio de domicilio, etc.

Este caso puede verse como una generalización del caso de censura por la derecha explicado anteriormente. Ahora la variable C_r no es una cantidad de tiempo fijada de antemano por el investigador, sino que es una variable aleatoria temporal asociada a otro suceso ajeno e independiente del suceso final en el que estamos interesados. Asumimos que C_r y T son variables independientes.

3.4.5. Censura en intervalos

Es el tipo de censura que ocurre cuando un tiempo de vida solo es conocido durante un intervalo de tiempo. Esto puede ocurrir, por ejemplo, en experimentos industriales donde los períodos de inspección son periódicos. Podemos decir que este tipo de censura es una generalización de la censura por derecha y por la izquierda.

3.5. Tipos de Truncamiento

El truncamiento es una condición que presentan ciertos sujetos en el estudio y que el investigador no puede considerar su existencia. Se definen:

3.5.1. Truncamiento por la izquierda

Ocurre cuando los sujetos entran en el estudio en un momento en particular (no necesariamente el origen del evento de interés) y son seguidos desde ese momento de entrada tardío hasta que ocurre el evento o hasta que el dato es censurado.

3.5.2. Truncamiento por la derecha

Ocurre cuando solo individuos que han presentado el evento son incluidos en el estudio.

4. Metodología Estadística (ME)

4.1. Introducción a la estimación por núcleos

Muchas veces cuando queremos estimar una densidad en Estadística (o en este caso más concreto, una función de Riesgo), recurrimos al clásico método del Histograma, ya que es el método más intuitivo de aplicar. Básicamente lo que hacemos, suponiendo un caso de una dimensión, es elegir un origen x_0, del cuál se parte a estudiar los datos y luego dividir el eje de X (la variable en estudio) en particiones, generalmente llamadas *bins*, que pueden ser de un ancho h fijo o variable. Luego se ve para cada partición, cuantas observaciones caen dentro del mismo y así se va logrando el histograma. Por lo tanto para cada $x \in B_j$ se le asigna el valor:

$$\hat{f}_h(x) = \frac{1}{nh} \cdot \sum_{j=1}^{n} \text{cant. de observaciones en } B_j \qquad \text{si } x \in B_j$$

El histograma depende así de 2 parámetros relevantes: el origen x_0 y el ancho de la ventana para cada intervalo h. Cualquier cambio en estos parámetros, puede devenir en histogramas totalmente diferentes.

Una forma de solucionar el tema del origen x_0, propuesto por Scott (Scott 1985) consiste en hacer un promedio de varios histogramas, con distintos orígenes y un mismo ancho de ventana h, lo que se conoce como *Averaged Shifted Histogram* (ASH). De esta forma logramos obtener una versión más suavizada del histograma original. Ambos estimadores se pueden ver con mayor detalle en el trabajo de (Bourel 2013).

Sin embargo una restricción que puede presentar el histograma es que a cada x perteneciente a a un intervalo B_j se le asigna el mismo valor estimado de $\hat{f}_h(x)$, lo que hace que que el histograma no sea una función continua (presenta saltos) y por lo tanto no sea derivable en estos puntos y tenga derivada nula en los puntos restantes, lo cual no es de utilidad si queremos justamente estimar una función continua.

Una alternativa a este método, es el uso del **estimador por núcleo** (o *Kernel Density Estimator*). En este caso en vez de particionar el eje en *bins* de ancho h, con determinado origen y contar cuantas observaciones caen en cada intervalo, Rosenblatt (Rosenblatt 1956) propone considerar para cada x, un intervalo centrado en x con radio h y ponderar, dando mayor peso a las observaciones que se encuentran cerca de x en $(x-h, x+h]$. El estimador por núcleo se define de la siguiente forma:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h\left(x - x_i\right) \tag{4}$$

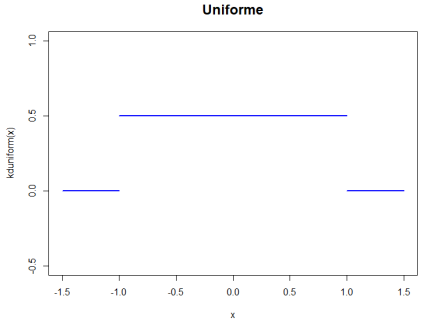
A K se la conoce como función núcleo y por lo general se pide que sea no negativa, simétrica e integre uno, por lo tanto cumple:

$$\int K(y)dy = 1$$

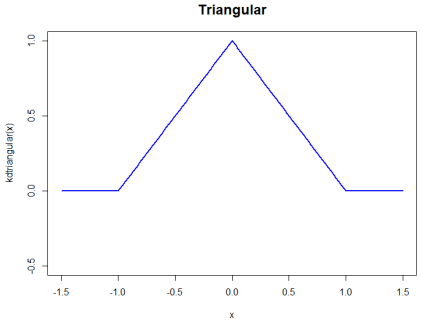
$$\int yK(y)dy = 0$$

$$\int y^2K(y)dy = \mu_2(K) > 0$$

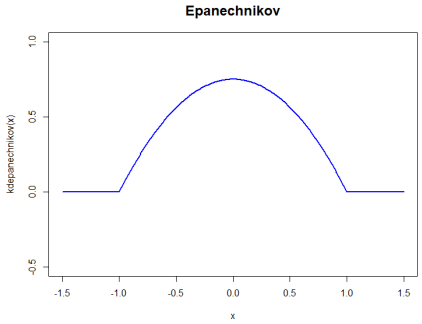
La función núcleo K indicará el peso que tiene la observación x_i en la estimación de $f(x)$. Son variados los ejemplos de funciones núcleo, en la Figura 1 se presentan algunos ejemplos.



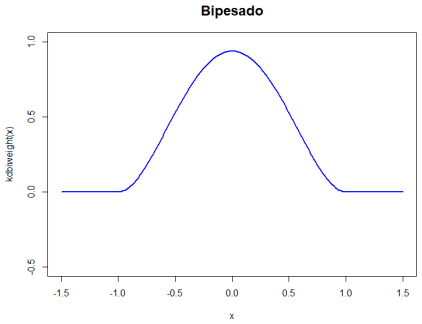
$K(x) = \frac{1}{2}I_{[-1,1]}(x)$



$K(x) = (1 - |x|)I_{[-1,1]}(x)$



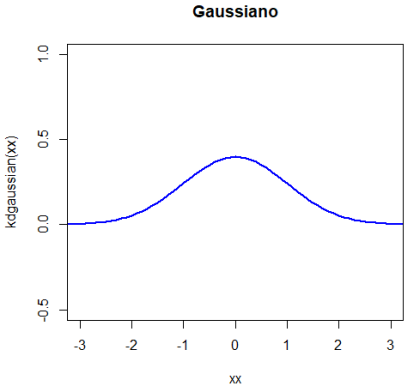
$K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}(x)$



$K(x) = \frac{15}{16}(1 - x^2)^2I_{[-1,1]}(x)$

Figura 1: Ejemplos de funciones de núcleos con soporte acotado.

Las funciones núcleo presentadas en la Figura 1 se conocen como núcleos de soporte acotado, es decir, valen lo que vale la función dentro de un soporte [-1,1]. Otro núcleo también muy utilizado comúnmente, que no es de soporte acotado, es el núcleo gaussiano:



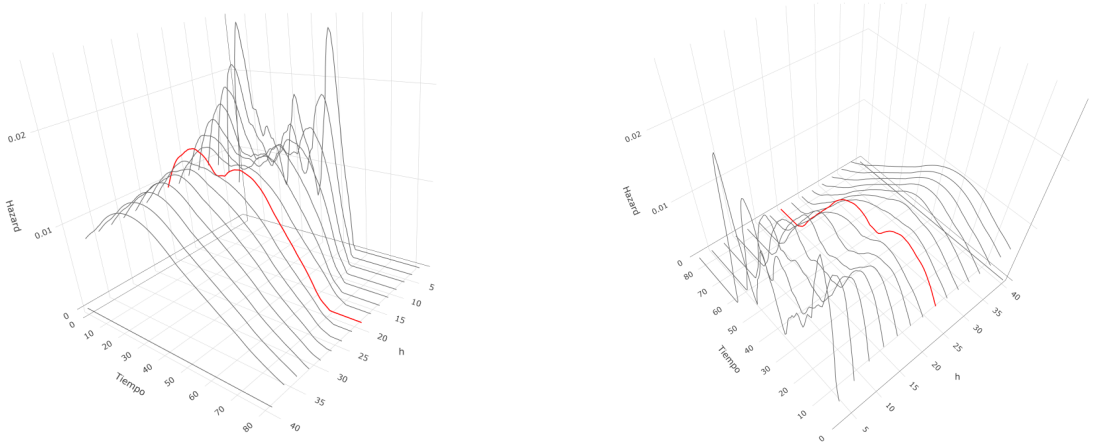
$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Figura 2: Ejemplo de función núcleo con soporte no acotado.

En el caso del núcleo uniforme todas las observaciones que caen en el intervalo $(x-h, x+h]$ tienen el mismo peso, por lo que se asemeja al estimador que se obtiene con un histograma de ventana móvil. Esto no pasa haciendo uso del núcleo triangular, el de Epanechnikov, el bipesado o el gaussiano, ya que las observaciones tienen distinto peso, siendo mayor cuanto mas cercanas se encuentren a x.

En este caso la estimación depende de tanto de la función núcleo K, como también al igual que los métodos anteriores, del ancho de ventana h. La elección del núcleo le da "forma" a la función y no es tan determinante como el ancho de la ventana, donde la elección de este último va a decretar que tan "suave" es la función.

Si se toma un h grande, más observaciones van a ser tenidas en cuenta en cada intervalo y por lo tanto la estimación será mas "suave" (generando más sesgo), en cambio si se elige un h pequeño, la función va a tomar en cuenta menos observaciones (o puede que ninguna) en cada intervalo y la función a estimar se va a volver más "rugosa" (generando más variabilidad). El tema del ancho de ventana óptimo es uno de los problemas que ocupa la estimación por núcleos, ya que se genera ese intercambio entre sesgo y varianza. Más adelante mencionaremos 2 métodos para la elección del ancho de ventana, particularmente para el caso que nos ocupa, la función de riesgo. En la Figura 3 se puede apreciar mejor este efecto en la elección.



a) Elección del ancho de ventana (óptimo en rojo)

b) Elección del ancho de ventana (otra vista)

Figura 3: Efecto de la elección del ancho de ventana sobre la estimación, en este caso de la función de riesgo univariada. Se ve que a medida que el h se hace más pequeño, mas variable se hace la estimación. Por el contrario, con valores de h grande se "alisa" más la función.

4.2. Estimación de la Función de Riesgo

4.2.1. Función de Riesgo Univariada

Como se mencionaba anteriormente, son varias las distribuciones de probabilidad que suelen utilizarse (Exponencial, Weibull, Gamma, Lognormal) y si conocemos específicamente la distribución de T, podemos estimar la función de riesgo de forma más precisa. Pero en muchos casos, esta información no la conocemos, por eso utilizamos métodos no paramétricos, en especial los estimadores por núcleo, que poseen buenas propiedades estadísticas (Horová, Kolácek y Zelinka 2012).

Como se explica en (Selingerová, Doleželová et al. 2016), la idea de la estimación por núcleos es que el valor de la función, desconocida a cierto tiempo t, se puede estimar como un promedio ponderado de las observaciones conocidas en un entorno de t. La función núcleo K es la que juega el rol de ponderador, y el ancho de ventana h determina el tamaño del entorno.

Retomando los conceptos que se habían visto en (MT, página 4), consideraremos el modelo de censura por derecha. Como se comentaba anteriormente en (MT), la función de riesgo se define como la probabilidad de que un evento de interés ocurra en el próximo instante dado que no ocurrió anteriormente, es decir:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t | T \geq t\right)}{\Delta t} \tag{5}$$

Uno de lo estimadores propuestos por los autores, está conformado por los tiempos observados ordenados $Y_{(i)}$ y sus correspondientes indicadores de censura $\delta_{(i)}$ (que valía 1 cuando el evento era observado y 0 cuándo el dato era censurado), quedando de la forma:

$$\hat{\lambda}(t) = \frac{1}{h} \sum_{i=1}^{n} K\left(\frac{t - Y_{(i)}}{h}\right) \frac{\delta_{(i)}}{n - i + 1} \tag{6}$$

La función de estimación anterior, denominada por los autores como estimador *interno*, se presenta así como una convolución entre el estimador por núcleo y el estimador de la función de riesgo acumulada de Nelson-Aalen (el segundo término de la sumatoria). El estimador de Nelson-Aalen es un estimador no paramétrico que se utiliza por lo general en caso de tener datos censurados o incompletos. Se puede ver con mayor detalle en el Capítulo 2 de (Therneau y Grambsch 2000).

Previamente se había visto que hay una variable aleatoria T que representa el tiempo de supervivencia de un individuo con función de supervivencia $\overline{F}(\cdot)$ y densidad $f(\cdot)$. Por otro lado se encuentra la variable aleatoria censurada C_r, independiente de T con función de supervivencia $\overline{G}(\cdot)$ y densidad $g(\cdot)$. De esta forma se tiene una variable aleatoria no negativa Y, que representa el tiempo de supervivencia de un individuo, siendo $Y = min(T, C_r)$, con función de supervivencia $\overline{L}(\cdot)$ y densidad $l(\cdot)$. Se puede deducir de esta manera que:

$$\overline{L}(t) = P(Y \geq t) = P(T \geq t, C \geq t) = \overline{F}(t)\overline{G}(t) \tag{7}$$

Si llamamos $r(\cdot) = l(\cdot, 1)$ a la subdensidad de las observaciones sin censura (presentaron el evento), que puede ser expresado de la forma $r(t) = f(t)\overline{G}(t)$, podremos reescribir la función de riesgo de la ecuación (2) como:

$$\lambda(t) = \frac{f(t)}{\overline{F}(t)} = \frac{f(t)\overline{G}(t)}{\overline{L}(t)} = \frac{r(t)}{\overline{L}(t)} \tag{8}$$

Por lo que otra aproximación de la estimación se puede obtener mediante la función propuesta por (Selingerová, Horová y Zelinka 2014) y (Selingerová, Doleželová et al. 2016), denominada por los autores como estimador *externo*, y se define como la relación entre la densidad de las observaciones que presentaron el evento (muertes/fallos) y la función de supervivencia del tiempo observable T, sustituyendo $r(t)$ y $\overline{L}(t)$, por sus estimadores por núcleos:

$$\hat{\lambda}(t) = \frac{\frac{1}{h}\sum_{i=1}^{n} \delta_i K\left(\frac{t-T_i}{h}\right)}{\sum_{i=1}^{n} W\left(\frac{T_i-t}{h}\right)} \tag{9}$$

donde W es una función núcleo de distribución, tal que $W(y) = \int_{-\infty}^{y} K(t)dt$.

Así de esta manera (y como se mencionó anteriormente) podemos llegar también al estimador por núcleo de la función de supervivencia, haciendo uso de la relación que tiene con la función de riesgo. Habíamos visto que $H(x) = -log\overline{F}(x)$, por lo tanto quedaría de la forma:

$$\hat{\overline{F}}(t) = \exp^{-\int_0^t \hat{\lambda}(u)du} \tag{10}$$

4.2.2. Función de Riesgo Condicional

Estudiar la función de riesgo univariado es una primera aproximación al problema de interés, sin embargo en muchas ocasiones resulta poco informativo. Por lo general el riesgo, además de por el tiempo de seguimiento, se encuentra condicionado a ciertos atributos que puedan presentar los individuos o elementos de estudio. Por ejemplo, en casos clínicos o biomédicos, nos puede interesar estudiar el riesgo condicionado a la edad del paciente, sexo de la persona, avance de la enfermedad, tamaño del tumor (en caso de ciertos tipos de cáncer), o como puede ser también el tipo o densidad de un material, en casos de que estemos estudiando una pieza de maquinaria. Al condicionar por una covariable, la función de riesgo condicional se puede expresar de la siguiente manera:

$$\lambda(t|x) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t | T \geq t, X = x\right)}{\Delta t} \tag{11}$$

En este caso, y extendiendo lo visto para el caso univariado, el estimador interno por núcleo de la función de riesgo condicional que proponen los autores (Selingerová, Katina y Horova 2021) es:

$$\hat{\lambda}(t|x) = \frac{1}{h_t}\sum_{i=1}^{n} K\left(\frac{t - Y_{(i)}}{h_t}\right) \frac{\delta_{(i)} w_{(i)}(x)}{1 - \sum_{j=1}^{i-1} w_{(j)}(x)} \tag{12}$$

dónde $w_{(i)}(x)$ son los pesos de Nadaraya-Watson de la covariable analizada, que se definen como:

$$w_{(i)}(x) = \frac{K\left(\frac{x-X_i}{h_x}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_x}\right)} \quad \text{, con } i = 1, ..., n \tag{13}$$

Estos pesos son utilizados por lo general en regresión no paramétrica y van a actuar como ponderador en el sentido de la covariable.

Por otro lado también tenemos otra función propuesta tanto en (Selingerová, Horová y Zelinka 2014), como en (Selingerová, Doleželová et al. 2016), y al igual que para el caso univariado, se presenta como la relación entre el estimador núcleo de la densidad condicional de las observaciones sin censura y el estimador núcleo de la función de supervivencia condicional del tiempo observable T:

$$\hat{\lambda}(t|x) = \frac{\frac{1}{h_t}\sum_{i=1}^{n} w_i(x) K\left(\frac{t-T_i}{h_t}\right)\delta_i}{\sum_{i=1}^{n} w_i(x) W\left(\frac{T_i-t}{h_t}\right)} \tag{14}$$

donde diferenciamos W, que es una función núcleo de distribución, como mencionábamos anteriormente y $w_i(x)$, los pesos de la covariables de Nadaraya-Watson. Esta última función define el llamado estimador externo por núcleo de la función de riesgo condicional.

4.3. Métodos de elección de ancho de ventana

Como se mencionaba anteriormente en (ME), la elección del ancho de ventana es determinante a la hora de estimar una función mediante núcleos, debido al intercambio entre sesgo y varianza que se genera; un ancho de ventana muy grande puede sesgar mucho la estimación, mientras que un ancho de ventana demasiado pequeño puede hacerla muy variable.

Los siguientes métodos se explicarán para el caso univariado para simplificar su lectura, sin embargo se pueden extender al caso condicional, teniendo en cuenta la covariable en la función de riesgo estimada.

4.3.1. Validación Cruzada

Para llegar a un h_{opt} partimos de la definición del Error Cuadrático Integrado (ISE) (Anexo 9.1)

$$ISE(\hat{\lambda}(\cdot,h)) = \int_0^S \left(\hat{\lambda}(x,h) - \lambda(x)\right)^2 dx, \quad \text{(desarrollando el cuadrado de binomio)} \tag{15}$$

$$ISE(\hat{\lambda}(\cdot,h)) = \int_0^S \hat{\lambda}^2(x,h) - 2\int_0^S \frac{\hat{\lambda}(x,h)}{1-F(x)} f(x)dx + \int_0^S \lambda^2(x)dx \tag{16}$$

El tercer término no depende de h, por lo que minimizar el valor por validación cruzada (CV) equivale a minimizar los dos primeros términos y a su vez el estimador insesgado del segundo término toma la forma:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\lambda}_{-i}(X_i,h)}{1-L_n(X_i)}\delta_i$$

Dónde $\hat{\lambda}_{-i}(X_i,h)$ es la estimación de la función de riesgo en X_i omitiendo ese punto. Por lo tanto la función de validación cruzada queda como:

$$CV(h) = \int_0^S \hat{\lambda}^2(x,h)dx - \frac{2}{n}\sum_{i=1}^{n} \frac{\hat{\lambda}_{-i}(X_i,h)}{1-L_n(X_i)}\delta_i \tag{17}$$

y la elección de un ancho de ventana óptimo a través de este método es aquel que minimiza la

función anterior:

$$\hat{h}_{CV} = argmin_{h \in H_n} CV(h) \tag{18}$$

En (Selingerová, Horová y Zelinka 2014, Sec. 3.2) podemos ver este método extendido para el caso de la estimación condicional, con un mayor grado de detalle.

4.3.2. Máxima Verosimilitud

El método modificado de verosimilitud fue propuesto por (Tanner y Wong 1983). La idea del método es elegir un h que maximice la función modificada de verosimilitud:

$$ML(h) = \prod_{i=1}^{n} \hat{\lambda}_{-i}^{\delta_i}(X_i, h)\overline{F}_{-i}(X_i) \tag{19}$$

donde $\hat{\lambda}_{-i}^{\delta_i}$ es el mismo que en el método de validación cruzada y $\overline{F}_{-i}(x) = exp^{-\int_0^x \hat{\lambda}_{-i}(t,h)dt}$.
El ancho de ventana óptimo a través de este método es entonces:

$$\hat{h}_{ML} = argmax_{h \in H_n} ML(h) \tag{20}$$

4.4. Puntos de cambio rápido

Hay ciertos puntos que son de interés, dónde se da la disminución más rápida de la función de riesgo, los cuales se denominan "puntos de cambio rápido" (*'points of the most rapid change'*). Estos puntos se dan en el extremo de la primera derivada de λ y se pueden detectar como ceros en la estimación de la segunda derivada de $\hat{\lambda}$. Solamente se toman como relevantes los ceros en esta función, cuando el signo pasa de ser negativo a positivo, dado que solo el mínimo local de $\hat{\lambda}$ es importante.

Para hallar la segunda derivada de la función de riesgo, lo que debemos es utilizar la derivada segunda del núcleo y ajustar la fórmula como se puede ver en (Horová, Kolácek y Zelinka 2012, pag. 120). En este caso, a modo de simplificación, podemos ver para el caso univariado, que esta función queda de la forma:

$$\hat{\lambda}(t) = \frac{1}{h^3} \sum_{i=1}^{n} K^{(2)} \left(\frac{t - Y_{(i)}}{h} \right) \frac{\delta_{(i)}}{n - i + 1} \tag{21}$$

siendo $K^{(2)}$ la derivada segunda de la función núcleo. Para hallar ceros en esta función, debemos hacer uso de algún método de aproximación numérica, como puede ser el *método de la secante.* [1]

[1] Propuesto en (Horová, Kolácek y Zelinka 2012). Se pueden utilizar otros métodos para hallar ceros de una función.

4.5. Modelo de riesgos proporcionales de Cox

Como se mencionaba anteriormente, cuando se realiza un análisis de supervivencia (o en este caso del riesgo) lo que interesa conocer, además del tiempo en que puede ocurrir un evento analizado, es saber como impactan una serie de características en la variable en estudio.

El modelo de riesgos proporcionales introducido por Cox (Cox 1972) es el método de regresión multivariado más comúnmente utilizado en la investigación médica para analizar la asociación entre el tiempo de supervivencia de los pacientes y una o más variables predictoras.

El modelo de Cox se expresa mediante la función de riesgo denotada por $\lambda(t)$, lo cual expresa el riesgo de que se manifieste el evento al tiempo t, para el individuo i-ésimo. El modelo está definido por la siguiente función:

$$\lambda_i(t) = \lambda_0(t) exp\left(X_i(t)\beta\right) \tag{22}$$

dónde:

- $\lambda_i(t)$ es la variable dependiente; el riesgo al tiempo t considerado dadas las covariables X_j, $j = 1, ..., k$ para el individuo *i-ésimo.*
- X_j,$j = 1, ..., k$, las variables predictoras
- β_j, $j = 1, ..., k$,los coeficientes asociados a las k covariables, es una medida del impacto de la variable predictora
- $\lambda_0(t)$, el riesgo basal, corresponde al valor de la función de riesgo si todas las covariables x_k son iguales a cero.

Este método es utilizado frecuentemente en análisis de supervivencia como alternativa a otros como ser Kaplan-Meier o pruebas de rango logarítmico, debido a que estos son útiles cuando la variable predictora es categórica (e.g un tratamiento A frente a otro B), pero no para predictores cuantitativos como la edad o el peso por ejemplo. El modelo de Cox funciona tanto para variables predictoras categóricas como cuantitativas, además que extiende el análisis para evaluar simultáneamente el efecto de varios factores de riesgo en el tiempo de supervivencia.

Si bien este modelo es ampliamente utilizado en este tipo de análisis, se basa sobre ciertos supuestos para que la interpretación de los datos sea válida, donde esto no siempre se cumple. El supuesto que hace es que las covariables no varían en el tiempo (las trata como independientes del tiempo), lo cual es un supuesto fuerte y por lo tanto su principal limitación. Por ejemplo, variables predictoras como el peso o la edad de un paciente, son variables que sí varían en el tiempo, por lo que el análisis sería válido por un lapso de tiempo relativamente corto en estos casos.

Se dice que el modelo de Cox, es un modelo *semiparamétrico*, esto debido a que incluye una parte *parámetrica* y otra *no paramétrica*:

- La parte *paramétrica* se corresponde con $exp(\sum_{i=1}^{k} \beta_i X_i)$, conocida como *"risk score"*, donde β es el vector de parámetros de la regresión.
- La parte *no paramétrica* es la función de riesgo basal $\lambda_0(t)$, la cual es una función arbitraria, no especificada.

Al no tener especificada la función de riesgo basal se puede estimar los coeficientes de la regresión para varias situaciones. Por ejemplo, si el modelo paramétrico correcto para el análisis fuera el Weibull (uno de los modelos utilizados en este tipo de análisis), las curvas obtenidas con el modelo de Cox, serán similares a las del modelo Weibull. Claramente si supieramos efectivamente el modelo paramétrico correcto, usaríamos el mismo antes que el modelo de Cox, pero ante la duda de un error de especificación, el modelo de Cox se presenta como una buena opción.

Con respecto a la interpretación del modelo de Cox, no se hace directamente a través del coeficiente estimado β, sino de su exponencial $exp(\beta)$. Las cantidades $exp(\beta)$ son estimadores de la *razón de riesgo* (HR). Para variables dicotómicas por ejemplo, $exp(\beta)$ se interpreta como la cantidad de riesgo que se tiene con la presencia de la covariable en relación a la ausencia de la misma. En variables categóricas politómicas, $exp(\beta)$ se interpreta como la cantidad de riesgo con respecto a una categoría de referencia. En variables cuantitativas (numéricas), $exp(\beta)$ representa la razón de riesgo al incrementar en una unidad la covariable.

Una razón de riesgo mayor a 1 indica una covariable asociada positivamente a la probabilidad que ocurra el evento y por lo tanto asociado negativamente a la supervivencia. Esto lo podemos resumir de la siguiente forma:

- $HR < 1$: Reducción en el riesgo
- $HR > 1$: Incremento en el riesgo
- $HR = 1$: No tiene efecto sobre el riesgo

Dado que la razón de riesgo de 2 sujetos con covariables fijas X_i y X_j:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)e^{X_i\beta}}{\lambda_0(t)e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}} \tag{23}$$

es constante durante el tiempo (como se ve en la fórmula anterior), el modelo de Cox también es conocido como modelo de *riesgos proporcionales.*

La estimación del vector de parámetros β en el modelo, está basado en una función de verosimilitud parcial que fue introducida por Cox (Cox 1972). El método de verosimilitud parcial se diferencia del método ordinario en tanto que el método ordinario se basa en el producto de las verosimilitudes de todos los individuos o elementos de la muestra, mientras que el método parcial se basa en el producto de las verosimilitudes de todos los cambios (o eventos) ocurridos. Para datos con tiempo de fallos sin empates, la función de verosimilitud parcial se define, como se puede ver en (Therneau y Grambsch 2000), de la siguiente manera:

$$PL(\beta) = \prod_{i=1}^{n}\prod_{t\geq 0}\left[\frac{Y_i(t)r_i(\beta,t)}{\sum_j Y_j(t)r_j(\beta,t)}\right]^{dN_i(t)} \tag{24}$$

donde $r_i(\beta,t)$ es la función score del riesgo (*"risk score"*) para el sujeto i, y se define como $r_i(\beta,t) = e^{X_i(t)\beta}$. El logaritmo de la verosimilitud parcial se puede escribir como:

$$l(\beta) = \sum_{i=1}^{n}\int_0^{\infty}\left[Y_i(t)X_i(t)\beta - log\left(\sum_j Y_j(t)r_j(t)\right)\right]dN_i(t) \tag{25}$$

Si bien no es una verosimilitud en el sentido estricto, se puede tratar como si lo fuera a efectos de hacer inferencia asintótica.

Si se diferencia el logaritmo de la verosimilitud parcial con respecto a β, obtenemos el *vector score* $U(\beta)$:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} [X_i(s) - \hat{x}(\beta, s)]\, dN_i(s) \tag{26}$$

donde $\hat{x}(\beta, s)$ es una media ponderada de X, sobre las observaciones que aún están en riesgo al tiempo s, esto se puede ver como:

$$\hat{x}(\beta, s) = \frac{\sum Y_i(s) r_i(s) X_i(s)}{\sum Y_i(s) r_i(s)} \tag{27}$$

siendo $Y_i(s)r_i(s)$ los pesos.

La segunda derivada negativa es la matriz de información y se describe como:

$$I(\beta) = \sum_{i=1}^{n} \int_0^{\infty} V(\beta, s) dN_i(s) \tag{28}$$

dónde $V(\beta, s)$ es la varianza ponderada de X al tiempo s:

$$V(\beta, s) = \frac{\sum_i Y_i(s) r_i(s) \left[X_i(s) - \hat{x}(\beta, s)\right]' \left[X_i(s) - \hat{x}(\beta, s)\right]}{\sum Y_i(s) r_i(s)} \tag{29}$$

El estimador de máxima verosimiltud parcial, se encuentra resolviendo la siguiente ecuación:

$$U\hat{\beta} = 0$$

La solución $\hat{\beta}$ es consistente y asintóticamente normal, distribuida con media β (el verdadero valor del parámetro) y varianza $(\mathcal{E}I(\beta))^{-1}$, la inversa de la matriz de información esperada. Esta esperanza requiere del conocimiento de la distribución de los datos censurados, por lo que se termina utilizando $I^{-1}(\hat{\beta})$ como varianza de $\hat{\beta}$.

Comúnmente se utiliza el algoritmo de Newton-Rapshon para resolver la ecuación de verosimilitud parcial. Se inicia con un valor $\hat{\beta}^{(0)}$ (por lo general $\hat{\beta}^{(0)} = 0$) y se itera el siguiente cálculo:

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + I^{-1}\left(\hat{\beta}^{(n)}\right) U\left(\hat{\beta}^{(n)}\right) \tag{30}$$

hasta la convergencia, evaluada por la estabilidad en el logaritmo de la verosimilutd parcial, es decir $l(\hat{\beta}^{(n+1)}) \approx l(\hat{\beta}^{(n)})$.

Para poner a prueba el valor del parámetro β hallado a través de la estimación del modelo de Cox, se pueden usar también en la función de verosimilitud parcial los test de Wald, *score* y razón de verosimilitud. La hipótesis nula que se prueba es $H_0 : \beta = \beta^{(0)}$, es decir, que el valor del coeficiente sea cero (y por lo tanto la covariable no tiene influencia sobre la función de riesgo). A continuación se detallan cada uno de ellos:

- La prueba de *razón de verosimilitud* es igual a dos veces la diferencia de la log-verosimilutd parcial evaluada en $\hat{\beta}$ y en $\beta^{(0)}$, es decir, entre la estimación final y la inicial, de la siguiente forma
$$2\left(l(\hat{\beta}) - l(\beta^{(0)})\right) \tag{31}$$

- La prueba de Wald es:
$$\left(\hat{\beta} - \beta^{(0)}\right)' \hat{I} \left(\hat{\beta} - \beta^{(0)}\right) \tag{32}$$
dónde $\hat{I} = I(\hat{\beta})$ es la matriz de información estimada. Para una única variable esto es igual al estadístico z: $\hat{\beta}/se(\hat{\beta})$.

- El estadístico *score test* es:
$$U'\left(\beta^{(0)}\right) I\left(\beta^{(0)}\right)^{-1} U\left(\beta^{(0)}\right) \tag{33}$$
En este test se utiliza el gradiente (derivadas) del logaritmo de la verosimilitud parcial evaluada en la hipótesis nula.

Bajo la hipótesis nula $H_0 : \beta = \beta^{(0)}$, la distribución asintótica que sigue cada una de las pruebas, es chi-cuadrado con k grados de libertad. Asintóticamente son equivalentes, pero en muestras finitas pueden diferir.

En la herramienta, para poder estimar dicho parámetro se hizo uso de la función `coxph` del paquete `survival` de R (Therneau y Grambsch 2000). Una vez estimado el parámetro, se utiliza el mismo en la ecuación del Modelo de Cox, donde el riesgo basal es una estimación univariada del riesgo mediante núcleos (4.2.1). Esta estimación semiparámetrica, entre otras. se puede ver con mayor detalle en (Selingerová, Katina y Horova 2021).

5. Aplicación shiny (AS)

Como se mencionaba anteriormente, el objetivo primero del trabajo es poder crear una herramienta computacional, para llevar a cabo este tipo de análisis, de forma que esté a disposición del usuario en cualquier momento. Se puede acceder a la herramienta a través del link https://diegoarare.shinyapps.io/VIAR. A su vez, para asegurar su reproducibilidad, se dispuso el código utilizado en un repositorio público, alojado en https://gitlab.com/diegoarare/viar.

5.1. Paquetes y librerías utilizadas

Para poder construir esta herramienta, se hace especial uso del paquete `shiny` (Chang, Cheng et al. 2021) del lenguaje `R` (R Core Team 2020). `shiny` es un paquete para la creación de aplicaciones web interactivas, que se puede lograr sin tener conocimientos de HTML, CSS o JavaScript (Wickham y Safari 2021). Este paquete se basa en una lógica de programación reactiva. La idea detrás de la programación reactiva es especificar un marco de dependencias, para que cuando el usuario ingrese o modifique un input (entrada) en la UI, todas las salidas relacionadas con ese input, se actualicen de forma automática. Estas aplicaciones se pueden "correr" directamente desde `R` o alojarse en algún servidor (para tener un mayor acceso), como puede ser shinyapps.io, por mencionar alguno de ellos.

Otro paquete relevante y que en parte corre en el *backend* de la aplicación web (lo que el usuario no ve), es el paquete `kernhaz` (Selingerová 2018). El paquete `kernhaz` fue creado por Iveta Selingerová y produce estimaciones de núcleo para la función de riesgo para datos censurados por derecha. Esto fue utilizado en la aplicación más que nada para la parte de análisis condicional, dado que el algoritmo ya se encontraba optimizado, dejando el caso univariado para trabajarlo de forma manual y poder conseguir otra información de interés como ser la función de supervivencia a través de núcleos y los puntos de cambio rápido mencionados anteriormente. Se puede conocer más del paquete aquí.

Además de estos paquetes se utilizan otros para añadir funcionalidades y visualizaciones. A continuación se presenta un breve resumen de los paquetes utilizados:

- **Para crear el entorno de la herramienta:**
  - `shiny` (Chang, Cheng et al. 2021)
    Como se mencionaba anteriormente, este paquete es utilizado para crear tanto la UI como el server, es decir, tanto el *frontend* (lo que el usuario ve y con lo que interactúa) y el *backend* (lo que corre detrás de la aplicación).
  - `shinydashboard` (Chang y Borges Ribeiro 2021)
    Si bien `shinydashboard` se puede utilizar para crear dashboards (tableros de gestión de información) basado en `shiny`, en este caso es utilizó para traer elementos (como los *box*) para mejorar el diseño de la herramienta.
  - `shinyWidgets` (Perrier, Meyer y Granjon 2022)
    Sirve para personalizar los elementos con los que el usuario interactúa (botones, sliders, listas, etc.)
  - `shinycssloaders` (Sali y Attali 2020)
    Es importantante que el usuario sepa cuando la herramienta está trabajando, por lo que en este paquete hay varios spinners para mostrar mientras se recalcula algún resultado.

- **Para realizar las estimaciones:**
  - `kernhaz` (Selingerová 2018)
    Se utilizó para realizar la estimación de la función de riesgo condicional principalmente y también para hallar anchos de ventana óptimos mediante los métodos mencionados anteriormente.
  - `survival` (Therneau y Grambsch 2000)
    Para realizar las estimaciones de los coeficientes del modelo de Cox, a través de la función `coxph`. También se utilizó para hallar la función de supervivencia de Kaplan-Meier, en el panel de riesgo univariado.
  - Funciones manuales.
    Se escribieron funciones manuales para lo que escapaba de los paquetes anteriores, en este caso por ejemplo para hallar los puntos de cambio rápido, para aplicar un modelo de Cox utilizando una función núcleo como riesgo basal y la estimación de la función de riesgo condicional para variables categóricas. Estas funciones se pueden ver en en el repositorio público, haciendo click en este link.
- **Para visualizar los resultados:**
  - `plotly` (Sievert 2020)
    Esta librería sirve para crear gráficos interactivos, con el fin de poder navegar de forma más fácil a través de los datos. El usuario interactúa con el gráfico, realizando cambios en los inputs y viendo como impacta en el resultado dicho cambio. De esta manera puede localizar y extraer información que en un gráfico estático quizás no pueda apreciar, por lo que es una herramienta muy útil al momento del análisis.
  - `ggplot2` (Wickham 2016)
    Es un paquete para crear gráficos de alta calidad, con una lógica de creación "por capas". En lugar de crear gráficos predeterminados, a través de distintas capas se puede adaptar el gráfico a cualquier problema. Se puede conocer más del paquete aquí.
  - `DT` (Xie, Cheng y Tan 2021)
    Se utiliza para mostrar los datos en formato de tablas en el panel de exploración, a las que se le puede agregar funcionalidades de filtro, búsqueda y clasificación de datos.

En las siguientes secciones se explicará cada una de las partes de la aplicación, a modo de manual.

5.2. Pantalla de inicio

Figura 4: Pantalla de inicio de VIAR

La Figura 4 muestra la pantalla de inicio de la aplicación, es el primer contacto con la misma. Allí podemos ver una breve descripción de la herramienta y en que consiste el análisis de supervivencia, para después pasar al tema principal que nos ocupa, la función de riesgo.

A partir de esta pestaña inicial, podemos movernos a la distintas secciones de la aplicación. En (2) vamos a la sección donde se puede cargar la base de datos de supervivencia que dispongamos, en (3) se encuentra la parte de análisis, tanto univariado como condicional, además de una sección dedicada al cálculo del ancho de ventana óptimo a través de las técnicas mencionadas anteriormente (validación cruzada y máxima verosimilitud) y podemos ir a (4) para un análisis con un enfoque más descriptivo de los datos cargados, para poder explorarlos desde allí. Siempre podemos volver a la pantalla de inicio a través de (1).

5.3. Carga de datos

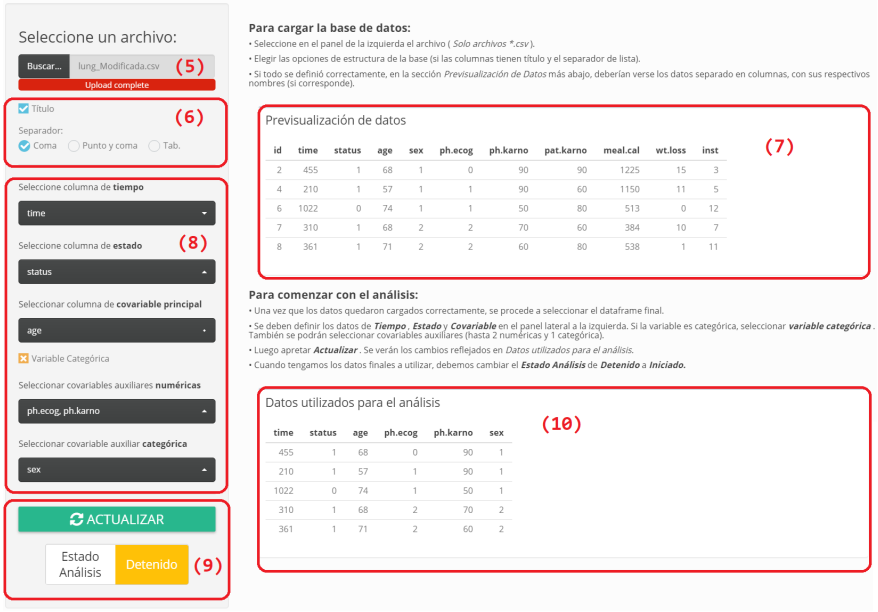


Figura 5: Panel de carga de datos

En este panel (Figura 5) se podrá cargar la base de datos que deseamos analizar, actualmente toma únicamente archivos de valores separados por comas (*.csv) como input, quizás más adelante se podrán adaptar otro tipo de extensiones. El mecanismo de carga consta de los siguientes pasos, que hay que seguir de forma secuencial para que quede todo correctamente cargado:

1. Cargamos el archivo que queremos analizar en (5), lo cuál apenas cargue se podrán ver los datos reflejados en (7), en formato de tabla.

2. En (6) tenemos varias opciones para modificar la estructura de datos cargados, es decir si las columnas tienen título o no, y cual es el separador de los datos. Esto último es importante, dado que si no ingresamos correctamente el separador, los datos van a quedar ingresados únicamente en una columna, lo cual es incorrecto. La forma correcta es que cada variable o campo ocupe una columna como se puede apreciar en la imagen. Todos los cambios que hagamos aquí se ven reflejados en (7).

3. En (8) debemos seleccionar, de todos los datos, los específicos de cualquier análisis de supervivencia, es decir, un tiempo hasta que ocurre un evento o no, el estado (si ocurre el evento o si es censurado) y covariables de interés para extender el análisis. La aplicación acepta hasta cuatro covariables, siendo una de ellas la principal, con la que se puede interactuar cambiando los valores de ancho de ventana, núcleo y tipo de estimador y otras tres auxiliares (dos numéricas y una categórica), que tienen un ancho de ventana fijo (en todo caso se pueden cambiar al gráfico principal para profundizar en su análisis). La covariable principal, si es categórica, es necesario seleccionar la opción **"Variable categórica"**, ya que si no se realiza esto, puede dar visualizaciones erróneas.

   Otro tema relevante a comentar en este punto, es que el **estado solo toma valores binarios 0-1**, es decir:

   - Valor 0: Dato censurado
   - Valor 1: Dato con presencia del evento

   , cualquier otro valor de entrada puede generar errores en el algoritmo.

4. Una vez que ingresamos todo correctamente en el paso 3, le damos al botón "Actualizar" en (9) y podemos visualizar este nuevo dataframe en (10).

5. Si se ven los datos de forma correcta y no deseamos hacer más ajustes, cambiamos en (9), el *Estado Análisis* de *Detenido* a *Iniciado* y estamos prontos para comenzar.

Es importante aclarar, que los datos ingresados deben ser todos númericos, incluso aquellos que representan variables categóricas (por ejemplo usar una variable binaria 0-1, para datos como el sexo de la persona).

5.4. Exploración de Datos

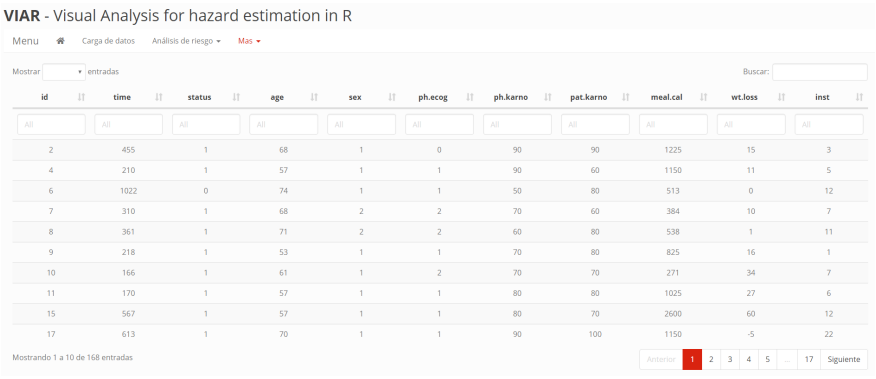


Figura 6: Tabla de datos

En la sección *Más*, tenemos opciones para poder explorar los datos cargados. Por un lado tenemos en la pestaña de *"Datos"* (Figura 6), una tabla donde observar, buscar y filtrar los datos cargados. Cabe destacar que esta reúne toda la base cargada, no únicamente los datos utilizados para el análisis. Cada columna tiene un buscador, en el cúal podremos buscar el número que se desee o incluso un rango determinado.

A través de la pestaña *"Descriptiva"*, se podrá acceder a información descriptiva más detallada de la base (con respecto a la que se veía en los paneles de análisis).

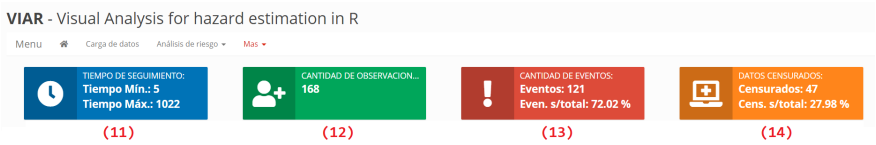


Figura 7: Información descriptiva de la base de datos

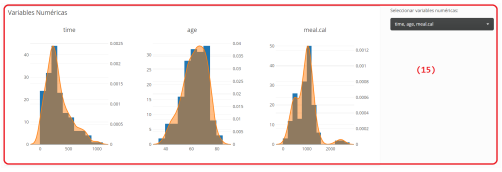


Figura 8: Gráficos de variables numéricas mediante 2 métodos: histograma (en azul) y densidad kernel (en naranja).

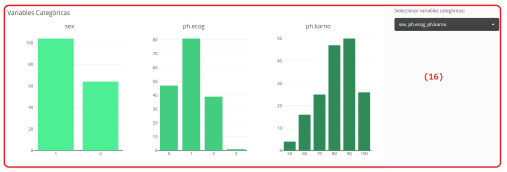


Figura 9: Gráficos de barra para variables categóricas

En la Figura 7, se aprecian varias *cajas* de información, donde se puede ver información sobre el tiempo de seguimiento del estudio (11), la cantidad de observaciones que disponemos en la base (12), la cantidad de eventos ocurridos y el porcentaje que representa en el total (13) y los datos censurados, es decir las observaciones donde el evento no se reflejó (14), también con su respectivo porcentaje sobre el total. En las Figuras 8 y 9, se tiene a disposición 2 secciones, una para variables numéricas (15), donde se puede observar la distribución mediante histogramas y densidad kernel

de las variables numéricas y otra para variables categóricas (16). Cada uno de ellas acepta hasta 3 variables a estudiar.

5.5. Análisis Univariado

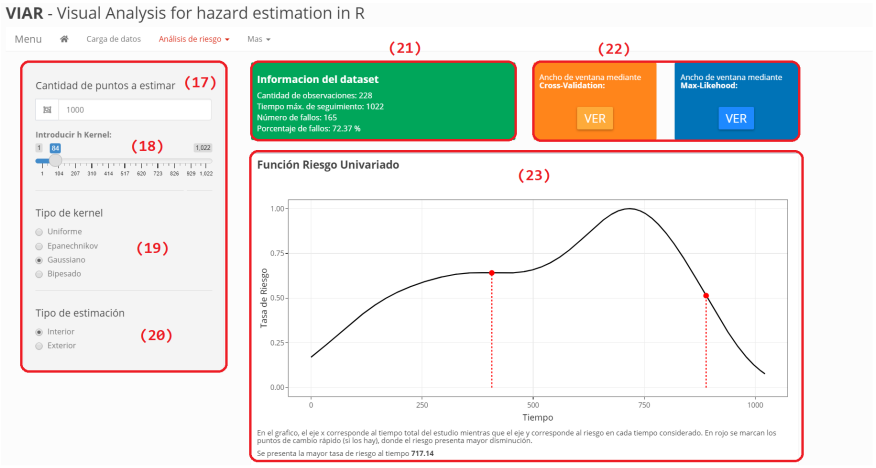


Figura 10: Panel Función de Riesgo Univariado

En este panel (Figura 10) comenzamos con el análisis. Como se mencionaba anteriormente, es un primer acercamiento a los datos, pero ya nos puede brindar información relevante como la forma que toma la función, la función de supervivencia, los puntos de cambio rápido y demás.

Contra la izquierda del panel, podemos ver todas las opciones (inputs) que podemos modificar. En (17) podemos elegir un número de puntos a estimar la función. Lo que hace esto básicamente es tomar el número ingresado y partir el rango de tiempo de los datos observados en esa cantidad de intervalos equidistantes, para luego poder correr el algoritmo. Cuanto menor sea el número, menos fina va a ser la estimación pero más rápida en el cálculo. Caso contrario cuanto mayor sea el número.

En (18) tenemos una de las partes relevantes de este tipo de técnica, el ancho de ventana. Como se mencionaba antes en el informe, el h es el valor determinante en el intercambio de sesgo-varianza que tiene la función. Un valor de h pequeño puede hacer la función muy *rugosa* (más variable), mientras que un valor muy grande la sesga mucho. La idea es poder encontrar cierto h óptimo.

En (19) podemos elegir 4 tipos distintos de núcleo: Uniforme, Epanechnikov, Gaussiano y Bipesado. Recordemos que estos cumplen la función de ponderador de las observaciones.

En (20) podemos seleccionar los 2 tipos de estimadores que se veían en 4.2.1 y 4.2.2. Para seguir con la terminología que utilizan los autores, se llamaron *Interno* y *Externo*. *Interno* hace referencia al estimador núcleo basado en la convolución de la función de núcleo, con un estimador no paramétrico de la función de riesgo acumulada. Por otro lado, *Externo* es el estimador definido como la relación entre el estimador núcleo de la subdensidad de las observaciones no censuradas y la función de supervivencia del tiempo observable.

Podemos observar en (21), de forma breve, información descriptiva del conjunto de datos, como ser la cantidad de observaciones, el tiempo máximo de seguimiento ingresado, la cantidad de eventos observados y su respectivo porcentaje sobre el total. Esto está disponible tanto en esta sección

como en la del análisis condiciona, de forma de no perder de vista esta información. En el panel de *Descriptiva* se podrá entrar más en detalle.

A través de (22) se podrá ir al panel de *Ancho de ventana* para intentar hallar un h óptimo. Esto lo veremos más adelante.

Finalmente en (23) se puede apreciar el gráfico de la función de riesgo univariado, dados todos los parámetros ingresados. Al utilizar el paquete `plotly` (Sievert 2020), este gráfico es interactivo, es decir podemos interactuar con él rotandolo para verlo desde distintos encuadres, ver los valores del riesgo en cada punto y acercarlo o alejarlo a través del zoom, entre otras opciones disponibles. Otros elementos que se pueden observar aquí son los puntos de cambio rápido (si los tiene) y dónde se halla el máximo valor de riesgo.

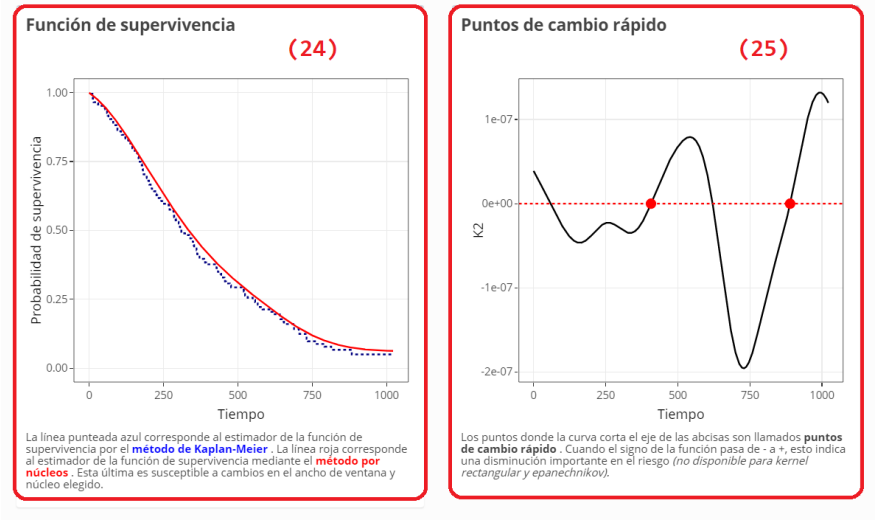


Figura 11:
Función de Supervivencia de Kaplan-Meier y a través de núcleos (24)
Puntos de cambio rápido a través de derivada segunda de la función de riesgo (25).

En la Figura 11 podemos observar, en contraposición a la función de riesgo, la función de supervivencia. En el gráfico se aprecia una aproximación a la misma a través de 2 técnicas diferentes: el método de Kaplan-Meier (línea punteada azul) y el método de núcleos (en rojo). Esta última es afectada por los parámetros de entrada (tipo de núcleo, ancho de ventana y tipo de estimación).

En (25) se puede ver los ceros en la derivada segunda de la función de riesgo, para ello se utilizó el método de la secante.

5.6. Análisis Condicional

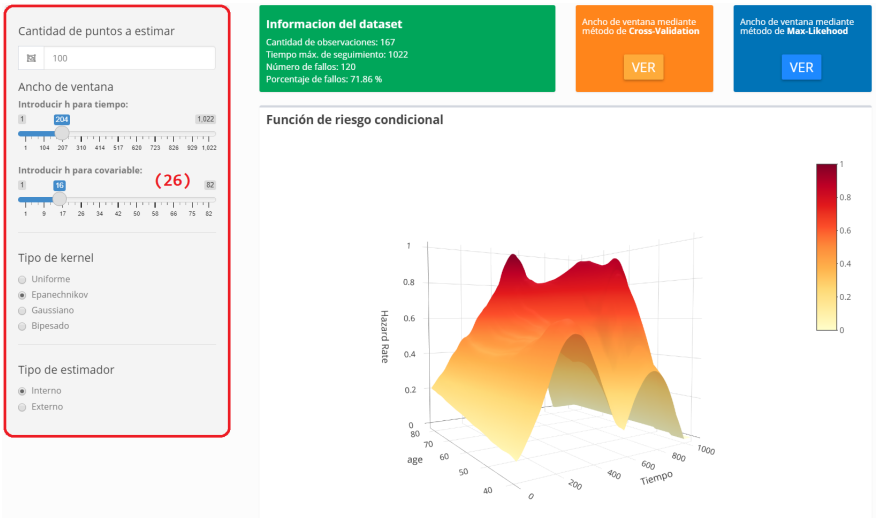


Figura 12: Panel de Análisis Condicional

En la Figura 12 se aprecia el panel del análisis condicional de la función de riesgo. Básicamente comparte los mismos parámetros que se veían anteriormente en el panel de análisis univariado. Lo que se agrega aquí es un *slider* más para el ancho de ventana de la covariable (26), así como teníamos del tiempo. Así vemos, que el entorno ya no está dado solo por el tiempo y por lo tanto ya no es un vector de datos (es decir un valor de riesgo asociado a cada tiempo), sino que el entorno está dado por una matriz de datos, dónde por ejemplo en los índices de las filas tendríamos los distintos tiempos, en los índices de las columnas los valores de la covariable y dentro de la matriz, los valores del riesgo asociado a cada uno de ellos.

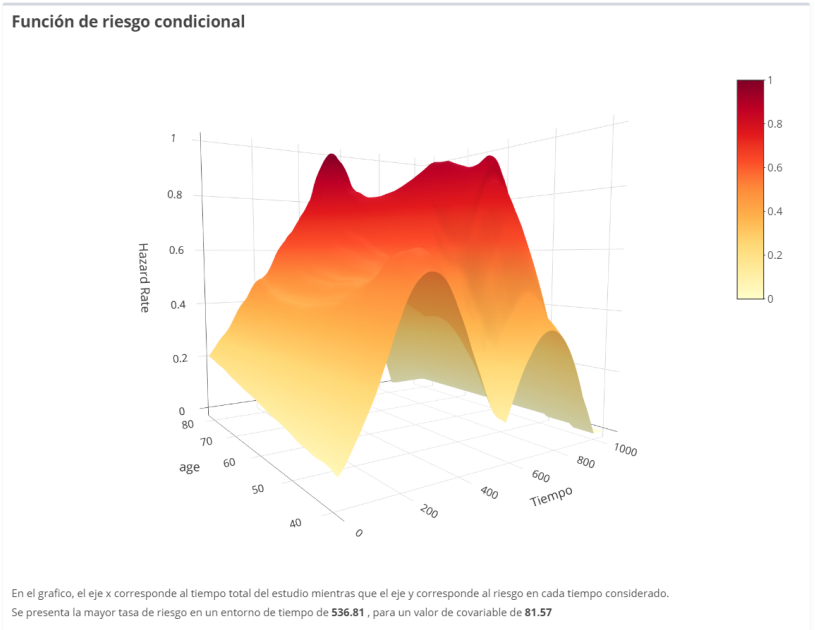


Figura 13: Gráfico de área de la función de riesgo rondicional. Las tasas de riesgo más altas se presentan en las zonas más oscuras.

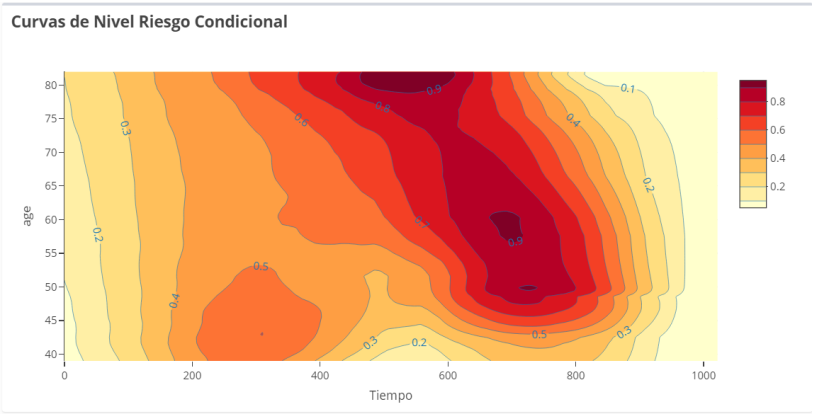


Figura 14: Gráfico de contorno de la función de riesgo condicional con curvas de nivel

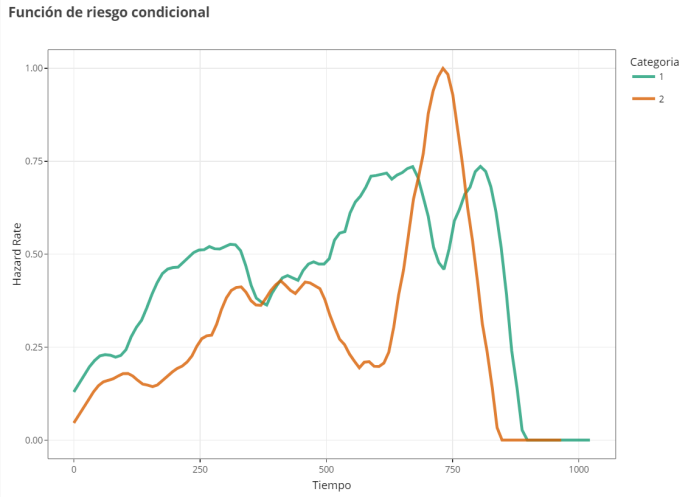


Figura 15: Gráfico de líneas de la función de riesgo condicional para variables categóricas. La altura determina las tasas más altas de riesgo.

En las Figuras 13 y 14, dados los parámetros de entrada que se hayan elegido, se puede ver el gráfico de la función de riesgo condicional desde distintas perspectivas para enriquecer el análisis. Este tipo de visualización se encuentra habilitado para covariables numéricas únicamente. En la Figura 15 se puede ver como queda la función de riesgo condicional para variables categóricas. Nuevamente el paquete `plotly` nos brinda mayor interactividad, para explorar el gráfico y poder aislar las zonas de mayor riesgo, para estudiarlas con mayor detalle.

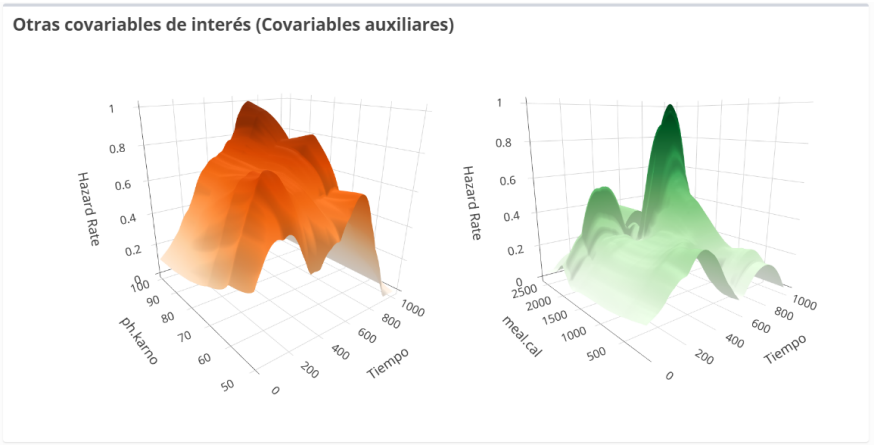


Figura 16: Gráficos de área de la función de riesgo condicional para distintas covariables auxiliares. Las mayores tasas de riesgo se presentan en las zonas más oscuras.

En la Figura 16, se puede observar una sección dedicada a otras covariables de interés. Como estudiamos de una covariable a la vez en este tipo de análisis, es necesario ver que pasa a su vez en otras covariables. Esto a modo de que si llegamos a detectar algo que llame la atención en estas covariables auxiliares, podemos ir nuevamente al panel de *Carga de datos* e intercambiar covariables y de esta manera verla en el gráfico principal (además de poder ajustar los parámetros de entrada en esta última opción).

5.7. Ancho de ventana óptimo

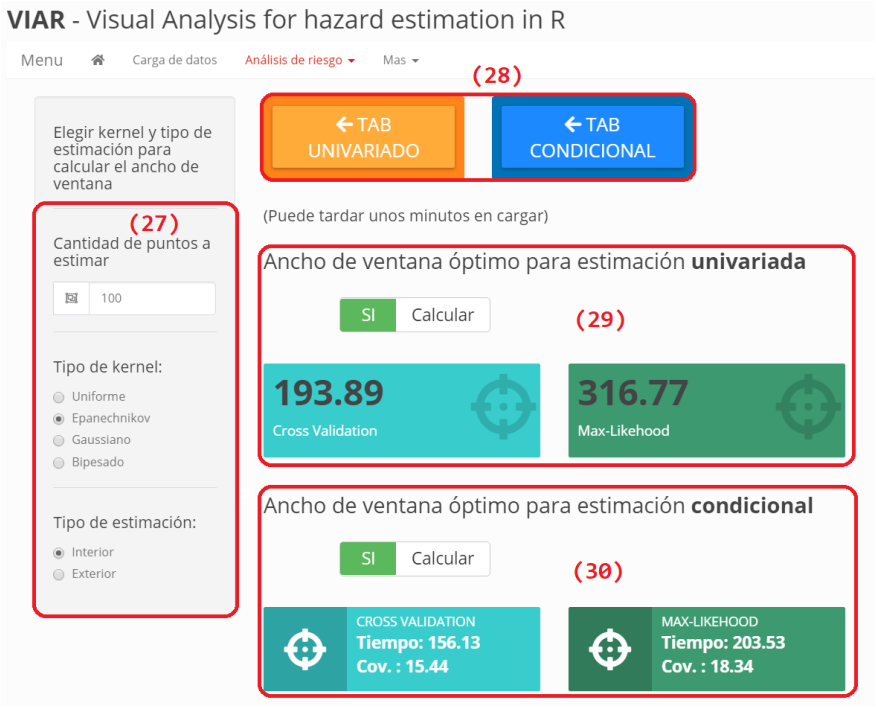


Figura 17: Panel Ancho de ventana óptimo

En este panel (Figura 17), podremos ver anchos de ventana sugeridos a través de las técnicas que se mencionaban anteriormente: por validación cruzada y máxima verosimilitud.

En (27) se encuentra la lista de opciones que veíamos tanto en el panel del análisis univariado como del condicional. En (29) y (30) se encuentran los módulos donde se corre el algoritmo que calcula el h óptimo para cada técnica y tipo de análisis. De forma predeterminada, se encuentra en estado suspendido. Para que inicie el cálculo hay que cambiar el estado del switch a `Calcular -> SI`. Esto puede demorar algunos minutos, dependiendo de la cantidad de datos que tengamos en la base.

Una vez hallados los h óptimos sugeridos por estas 2 técnicas, podremos volver a cualquiera de los 2 paneles de análisis, mediante (28).

5.8. Comparación Estimación Kernel con Modelo de Cox

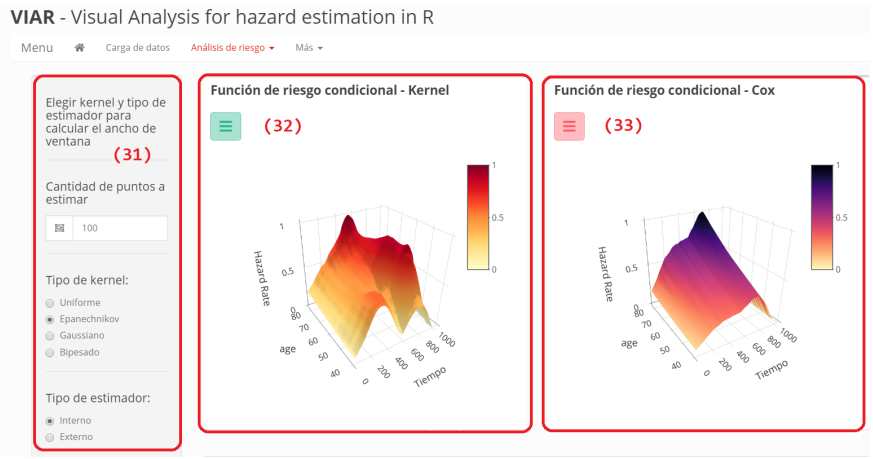


Figura 18: Comparación Kernel con Modelo de Cox

En esta sección (Figura 18) se podrá comparar el resultado obtenido con la metodología tratada, la estimación de la función de riesgo condicional mediante núcleos, con otro tipo de metodología generalmente utilizada, como es el Modelo de Cox. Como se mencionaba anteriormente, el Modelo de Cox es un modelo semiparamétrico, ya que tiene una parte paramétrica, correspondiente al *risk score*, $exp(\sum_{i=1}^{k} \beta_i X_i)$, y una parte no paramétrica que corresponde al riesgo basal. En este caso se utilizó un suavizado de la función de riesgo univariada mediante núcleos (como se puede ver en 4.2.1), para estimar el riesgo base.

En (31) podemos elegir las opciones para el suavizado mediante núcleos que afectan a ambas estimaciones, sin embargo en cada uno de los gráficos (32) para la estimación kernel y (33) para el Modelo de Cox, podremos seleccionar manualmente y de manera independiente el ancho de ventana. Esto debido a que mientras que en el Modelo de Cox hacemos una estimación univariada y por lo tanto solo debemos elegir un ancho de ventana para el tiempo, en la estimación de la función de riesgo condicional, el entorno queda determinado por el ancho de ventana del tiempo y de la covariable.

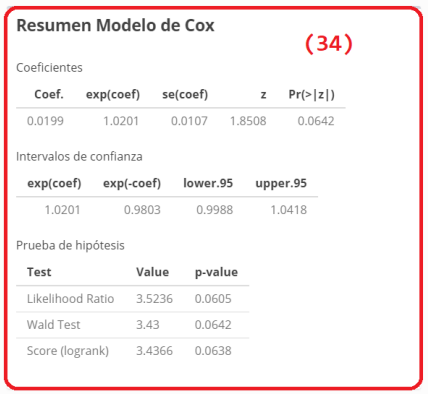


Figura 19: Resumen Modelo de Cox

En (34), de la Figura 19, se puede observar un resumen con la información del Modelo del Cox (*summary*), tanto del coeficiente β calculado como de las pruebas de hipótesis acerca del parámetro.

6. Ejemplo de uso para el Análisis: Melanoma (EJ)

6.1. Datos utilizados

Para realizar un análisis mediante la herramienta, se hace uso del dataset `melanoma` disponible en la librería `MASS` (Venables y Ripley 2002) de R. Este dataset surge de un estudio que se realiza en el período 1962-1977, e involucra a 205 pacientes diagnosticados con melanoma maligno, que fueron operados en el hospital universitario de Odense, en Dinamarca.

El melanoma es un tipo de cáncer de piel que se origina cuando los melanocitos (las células encargadas de producir *melanina* y que dan a la piel su color bronceado) comienzan a crecer fuera de control (Society s.f.). Estás células pueden convertirse en cáncer y extenderse a otras zonas del cuerpo. El melanoma es mucho menos frecuente que otros tipos de cáncer de piel (llamados genéricamente *cánceres de piel de tipo no melanoma*), pero muy peligroso porque es más probable que se propague a otras zonas del cuerpo (hace metástasis) si no se descubre y es tratado a tiempo.

Cuando una célula presenta alguna anomalía o ha envejecido, por lo general la célula muere. El cáncer (en general) se origina cuando algo no sigue dicho proceso, causando que las células anormales se reproduzcan y las células envejecidas no mueran, como debería suceder. Las células cancerosas pueden multiplicarse a números superiores que las células sanas y esto hace que el sistema del cuerpo no pueda funcionar correctamente.

Son varios los factores de riesgo para el cáncer de piel de tipo melanomas, algunos de ellos son:

- Exposición a la luz ultravioleta
- Presencia de lunares
- Poseer piel muy blanca, pecas y cabello claro
- Antecedentes familiares de melanoma
- Antecedente personal de melanoma
- Sistema inmunitario debilitado
- Envejecimiento

En el dataset que refiere al estudio, a cada paciente se le extirpó de forma completa el tumor mediante cirugía, además de extraer también 2.5 cm de la piel circundante. Entre las medidas que se tomaron se encuentra el grosor de tumor y si presentaba úlceras o no. Estas son variables prónosticas importantes en el sentido en que los pacientes con un tumor de mayor tamaño y/o ulcerado, presentan una mayor probabilidad de muerte por melanoma. El tiempo se presenta en meses desde el momento de la operación. A forma de resumen, se presentan las variables que contiene el dataset:

- **id**: Número de identificación del paciente.
- **time_months**: El tiempo medido en meses, desde el momento de la operación hasta el tiempo que finaliza el estudio o que el paciente fallece.
- **status**: Variable binaria, que indica si el paciente finaliza el estudio o sale del mismo por motivos ajenos a la enfermedad (0), o si fallece (1).

- **sex**: Variable binaria, si es femenino (0) o masculino (1).
- **age**: Edad del paciente al momento de la operación.
- **year**: Año en que fue realizada la operación.
- **thickness**: Espesor del tumor en mm. También se le llama *medición de Breslow*.
- **ulcer**: Si presenta ulceración (1) o no la presenta (0). La ulceración es la ruptura de piel que se encuentra sobre el melanoma.

6.2. Análisis y Resultados

Se empezará realizando un análisis exploratorio de las variables que se presentan en la base de datos, para luego poder estudiar algunas de ellas de manera más profunda a través de la función de riesgo. Es importante mencionar que el análisis siguiente responde a lo que está en este dataset únicamente, el cuál fue utilizado a modo de prueba de la herramienta. No se tuvo en consideración si es una muestra representativa de la población afectada por esta enfermedad.

6.2.1. Análisis exploratorio

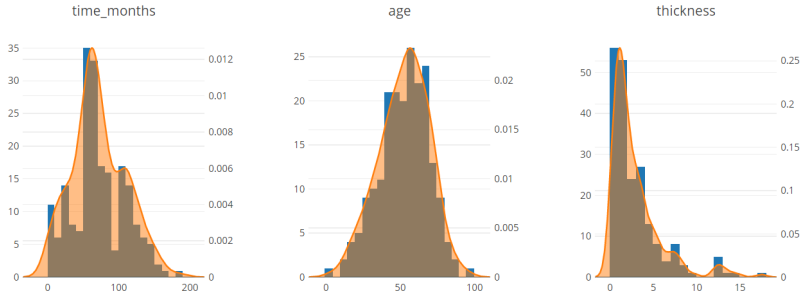


Figura 20: Análisis de algunas variables numéricas a través de histograma (azul) y densidad kernel (naranja). Las variables que se presentan de izquierda a derecha son el tiempo hasta el fallo o censura (*time_months*), edad del paciente (*age*) y espesor del tumor (*thickness*)

.

En la Figura 20 se puede observar que el tiempo en que se estudió la supervivencia del paciente luego de someterse a la cirugía va desde los 0.33 meses (aproximadamente 9 días) hasta los 185.5 meses (aproximadamente 15 años). Otras variables continuas de interés para el análisis son la edad y el espesor del tumor, debido a que ambas son factores de riesgo en la enfermedad, valores altos en estos dos campos aumentan el riesgo de fallecer por melanoma. Podemos ver que la edad del paciente va desde los 4 hasta 95 años, teniendo valor medio alrededor de los 52 años.

Con respecto al espesor del tumor (*tamaño*), visualmente ya se aprecia que por lo general se presentaron valores bajos. Casi una tercera parte del total de pacientes (56 pacientes, 27 % del total) presentan un tamaño de tumor menor a 1 mm y 173 pacientes (84 % aproximadamente) menores a 5 mm. Lo primero es importante, ya que la medida de melanomas con un grosor menor a un milímetro tienen una probabilidad muy pequeña de propagarse.

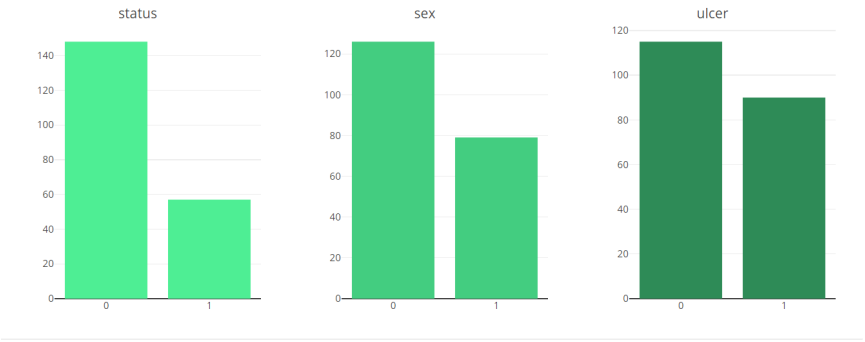


Figura 21: Análisis de algunas variables categóricas a través de gráficos de barras. Las variables presentadas son el estado (*status*, 0 para el dato censurado, 1 para el dato que presenta evento), sexo del paciente (*sex*, 0 indica sexo femenino y 1 sexo masculino) y si presenta ulceración o no (*ulcer*, 0 si no se presenta, 1 si presenta)

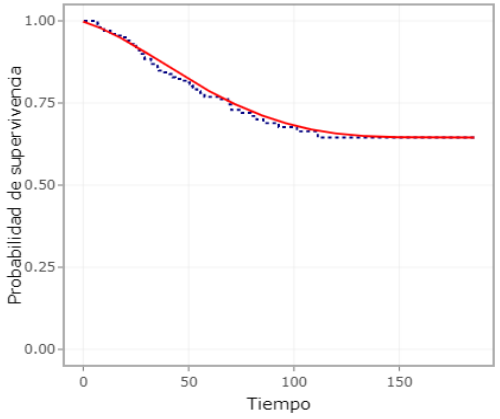
.

En la Figura 21 se puede ver la cantidad de eventos (fallecimientos) asociados al melanoma, en la variable *status*. Es preciso recordar que la categoría *0* hace referencia a que el dato es *censurado*, es decir, que llegaron al tiempo final de estudio sin haber presentado el evento o que dejaron el estudio por alguna otra razón (en esta categoría puede entrar también el fallecimiento por razones ajenas a la enfermedad de estudio). Se contabilizaron 57 pacientes fallecidos por melanoma (27,8 % del total) luego de haberse sometido a la extirpación del tumor.

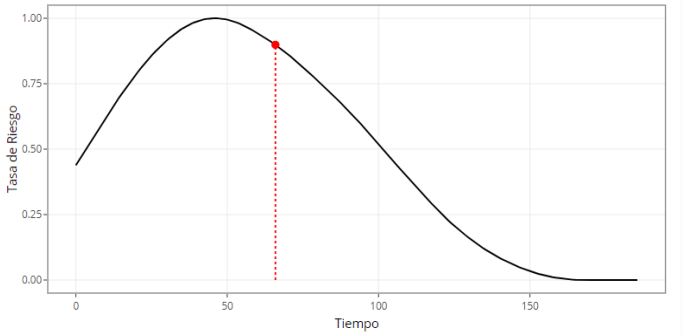
Otra variable relevante en el estudio de la enfermedad, como se mencionaba anteriormente, es la presencia o ausencia de ulceración sobre el melanoma. Es un factor de riesgo importante, ya que los melanomas ulcerados tienden a presentar un peor pronóstico. En este caso, no se presentan niveles tan dispares, siendo la mayoría pacientes que no presentan ulceración (115 pacientes no presentan úlceras contra 90 que presentan).

En cuanto al sexo de las personas en estudio, vemos una predominancia del sexo femenino sobre el masculino, una razón alrededor de 60-40.

6.2.2. Estimación de la función de riesgo univariada y condicional



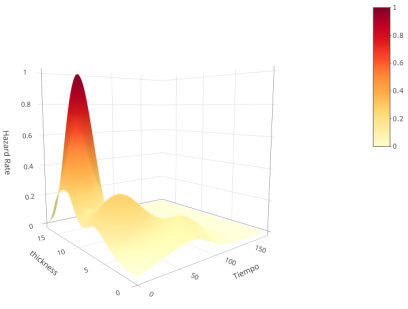
(a) Función de supervivencia, mediante Kaplan Meier (azul) y núcleos (rojo).



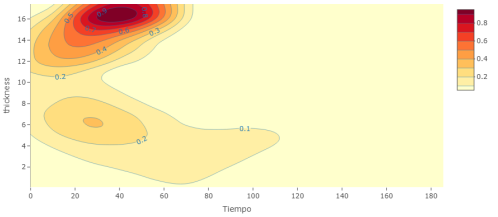
(b) Función de riesgo univariado mediante estimador por núcleos *interno*. En rojo se aprecia un punto de cambio rápido.

Figura 22: Función de supervivencia a través de 2 métodos (a) y función de riesgo univariado a través de la estimación por núcleos (b).

En la Figura 22.a se puede observar que la probabilidad de supervivencia para 3 y 5 años (medidas de tiempo generalmente utilizadas en este tipo de análisis), es de 87 % y 78 % respectivamente. En la Figura 22.b, se aprecia la función de riesgo univariado, es decir, la estimación que tiene en cuenta únicamente el tiempo hasta que ocurre el evento. En la misma se puede ver una curva creciente hasta un punto máximo ubicado en el entorno de los 46 meses (casi 4 años), para luego descender de forma gradual. También se detecta un punto de cambio rápido a los 66 meses (5 años y medio), donde el riesgo disminuye de forma más rápida.



(a)

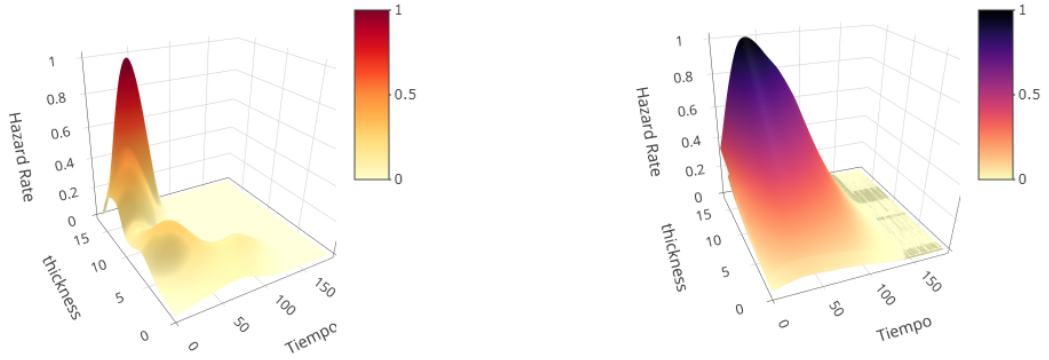


(b)

Figura 23: Estimación de la función de riesgo condicional con espesor del tumor como covariable a través de un gráfico de área (a) y gráfico de contorno con curvas de nivel (b). La zona más oscura indica un mayor riesgo.

En la Figuras 23.a y 23.b se observa la estimación del riesgo condicional, tomando como covariable una factor importante en este tipo de cáncer, el espesor del tumor. Tumores de mayor

tamaño son más complicados de tratar y eso se puede ver reflejado en ambos gráficos. En este caso vemos que el riesgo es importante para tumores con un espesor entre 15 y 17 mm, sobre todo en los primeros 30 a 45 meses. Sin embargo también se puede ver otro pico menos marcado, para tumores entre 5 y 7 mm.



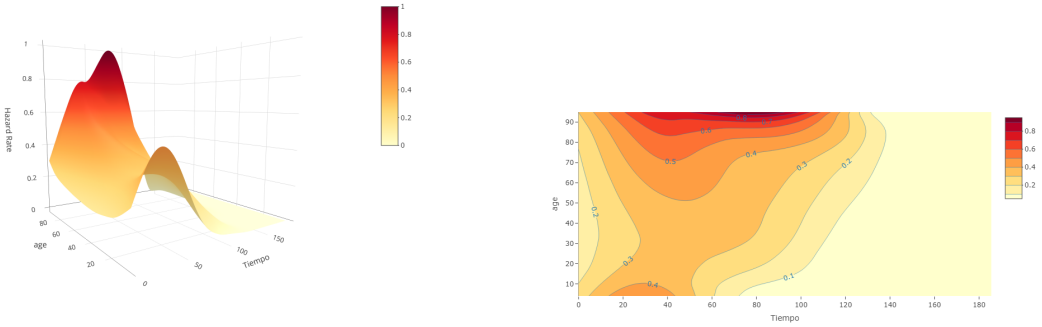
(a) (b)

Figura 24: Comparación de estimación de la función de riesgo condicional con *espesor de tumor* como covariable, mediante estimación por kernel (a) y a través del modelo de Cox (b)

En las Figuras 24.a y 24.b se puede ver la comparación de la estimación mediante núcleos, con el modelo de Cox ($\hat{\beta} = 0{,}16024$, $p - valor = 2{,}96e^{-07}$) para la covariable de *espesor del tumor*, dónde las diferencias entre un método y otro saltan a la vista. Mientras que el modelo de Cox solo puede capturar el riesgo en dirección al tamaño del tumor (a mayor tamaño, mayor riesgo), el método mediante núcleos puede detectar diferentes pendientes para distintos tiempos de supervivencia, lo cuál es una ventaja del método en este sentido.

Esto no quiere decir que el modelo de Cox no sea funcional, por el contrario, una de las principales ventajas del modelo de Cox sobre la estimación mediante núcleos, es la posibilidad de adaptar varias variables (y sus interacciones) a la estimación. En el caso de la estimación por núcleos, agregar más dimensiones, es sumamente complejo (computacionalmente), lo que la hace su principal desventaja.

Para seguir con el análisis, se puede observar el riesgo condicional a través de otro factor de riesgo que se mencionaba anteriormente, la edad. El melanoma es más probable que se presente en personas de edad avanzada, aunque también afecta a personas más jovenes. De hecho es uno de los tipos de cáncer más común en las personas menores a 30 años, especialmente en mujeres. El factor hereditario, es decir, el melanoma que se da por lo general entre personas de una misma familia puede manifestarse a una edad más temprana también.

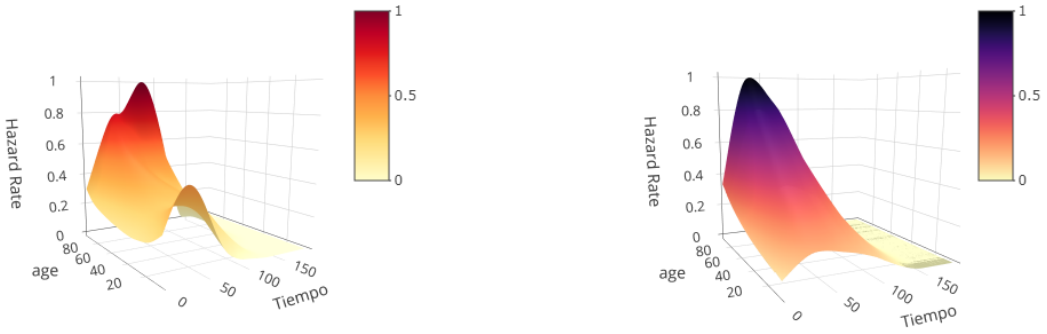


(a) (b)

Figura 25: Estimación de la función de riesgo condicional con *edad* como covariable a través de un gráfico de área (a) y gráfico de contorno con curvas de nivel (b). La zona más oscura indica un mayor riesgo.

En esta caso, en las Figuras 25.a y 25.b, se puede apreciar la función de riesgo condicional mediante núcleos. Nuevamente se puede ver distintas pendientes para los distintos tiempos de estudio y se refleja lo comentado anteriormente, se presenta mayor riesgo para las personas mayores y las más jovenes, representado por los dos picos que se ven en el gráfico. Se podría aislar (si la cantidad de observaciones lo permite) estos 2 grupos para estudiarlos de forma más profunda.

En la Figura 26.a y 26.b se realiza la comparación de la estimación mediante núcleos y el modelo de Cox ($(\hat{\beta} = 0{,}019220$, $p - valor = 0{,}0284)$) para la covariable *edad*. En el modelo de Cox se refleja que el método no puede captar esa edad temprana en la estimación del riesgo como factor relevante, únicamente lo hace para pacientes de edad avanzada.



(a) (b)

Figura 26: Comparación de estimación de la función de riesgo condicional con edad como covariable, mediante estimación por kernel (a) y a través del modelo de Cox (b)

A modo de resumen, se pueden destacar los siguientes puntos en cuanto a las metodologías utilizadas:

- Ambos modelos (Kernel y Cox) pueden ser válidos cuando los supuestos de los modelos paramétricos no se cumplen, ya que no hacen supuestos sobre la distribución del tiempo. Sin embargo, el modelo de Cox supone que las razones de riesgo son constantes a lo largo del tiempo, y la dependencia de la covariable es exponencial. Por lo que en este caso el método kernel puede ser muy útil, ya que no se hace supuestos.
- La estimación mediante núcleos (Kernel) puede capturar de mejor manera las distintas pendientes para cada tiempo y valor de covariable, algo que el modelo de Cox no logra destacar.
- Siguiendo el punto anterior, el método de Kernel logra una mejor visualización de los datos con respecto al modelo de Cox.
- Por otro lado, el método de Cox puede adaptar varias variables y sus interacciones al modelo, mientras que en el método Kernel agregar más covariables hace muy complejo el cálculo (además de su visualización).
- Otro punto a mencionar como desventaja de la estimación mediante núcleos es que la estimación es muy susceptible al ancho de ventana elegido, por lo que encontrar un valor óptimo es un tema central. En este trabajo se mencionaron 2 formas de poder aproximar un h óptimo, aunque pueden existir más.

Un punto relevante a tener en cuenta (que no es menor), es que el modelo de Cox que se utilizó como regla comparativa para evaluar la estimación por núcleos, es el modelo más simple (1 sola covariable). Como se mencionaba anteriormente, el modelo de Cox permite no sólo adaptar mas de una variable, sino que ver las interacciones entre ellas y devolver también una medida de resumen que expresa el efecto marginal de cada una. Por lo que la estimación por núcleos podría utilizarse para poder verificar los supuestos del modelo de Cox o para encontrar umbrales en las variables, principalmente en aquellas que son continuas (Selingerová, Doleželová et al. 2016).

7. Conclusiones y trabajo futuro

El objetivo principal del trabajo, era crear una herramienta interactiva para poder realizar el análisis de la función de riesgo univariada y condicional. La idea es que esto fuera un punto de partida para que la herramienta siga creciendo, pudiendo adaptarle más metodologías, no sólo para la función de riesgo, sino también para la función de supervivencia y así tener un análisis más completo.

La aplicación web quedó activa en https://diegoarare.shinyapps.io/VIAR/, donde se puede cargar la base de datos que se quiere analizar e interactuar con la herramienta. Se ha probado con varias bases, teniendo resultados satisfactorios y pudiendo realizar el análisis mencionado en el presente informe, haciendo uso solamente de la herramienta. A su vez, el código fuente quedó disponible en https://gitlab.com/diegoarare/viar, para asegurar su reproducibilidad.

Se ha podido adaptar la metodología propuesta por los autores de la Universidad de Mazaryk mencionados anteriormente y se ha hecho la comparación con otro tipo de método semiparamétrico como es el modelo de Cox, pudiendo ver las diferencias entre uno y otro.

Propuestas para trabajos futuros:

- Seguir agregando distintas metodologías a la aplicación, ya sean paramétricas (Weibull, Lognormal, etc.) o semiparamétricas (Modelo de Gray) para tener varias opciones a la hora de estimar.
- Seguir trabajando en la funcionalidad y el entorno "amigable" de la herramienta, para poder tener una mayor difusión y uso de la misma.
- Agregar opciones para automatizar el análisis, en el sentido de poder ingresar varios parámetros de entrada y la herramienta itere generando resultados que luego se impriman juntos en pantalla.
- Estudiar la manera de poder adaptar más covariables al método de estimación mediante núcleos.
- Optimizar los tiempos de cálculo a través de la paralelización, para mejorar el rendimiento.

8. Referencias Bibliográficas

Bourel, M. (2013). "Comparación en la elección de una ventana óptima para algunos estimadores de densidad". En: *Memoria Investigaciones en Ingeniería* 11, págs. 59-74. URL: http://revistas.um.edu.uy/index.php/ingenieria/article/view/357.

Chacón, J. y T. Duong (2018). *Multivariate Kernel Smoothing and its Applications*. Chapman y Hall/CRC. DOI: 10.1201/9780429485572.

Chang, W. y B. Borges Ribeiro (2021). *shinydashboard: Create Dashboards with 'Shiny'*. R package version 0.7.2. URL: https://CRAN.R-project.org/package=shinydashboard.

Chang, W., J. Cheng et al. (2021). *shiny: Web Application Framework for R*. R package version 1.6.0. URL: https://CRAN.R-project.org/package=shiny.

Cox, D. R. (1972). "Regression Models and Life-Tables". En: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2, págs. 187-220. ISSN: 00359246. URL: http://www.jstor.org/stable/2985181.

Horová, I., J. Kolácek y J. Zelinka (2012). *Kernel Smoothing in Matlab: Theory and Practice of Kernel Smoothing*. ISBN: 9814405485. DOI: 10.1142/8468.

Perrier, V., F. Meyer y D. Granjon (2022). *shinyWidgets: Custom Inputs Widgets for Shiny*. R package version 0.7.0. URL: https://CRAN.R-project.org/package=shinyWidgets.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rosenblatt, M. (1956). "Remarks on Some Nonparametric Estimates of a Density Function". En: *The Annals of Mathematical Statistics* 27.3, págs. 832-837. DOI: 10.1214/aoms/1177728190. URL: https://doi.org/10.1214/aoms/1177728190.

Sali, A. y D. Attali (2020). *shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating*. R package version 1.0.0. URL: https://CRAN.R-project.org/package=shinycssloaders.

Scott, D. W. (1985). "Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions". En: *The Annals of Statistics* 13.3, págs. 1024-1040. DOI: 10.1214/aos/1176349654. URL: https://doi.org/10.1214/aos/1176349654.

Selingerová, I. (2018). *kernhaz: Kernel Estimation of Hazard Function in Survival Analysis*. R package version 0.1.0. URL: https://CRAN.R-project.org/package=kernhaz.

Selingerová, I., H. Doleželová et al. (2016). "Survival of Patients with Primary Brain Tumors: Comparison of Two Statistical Approaches. PLOS ONE 11(2): e0148733". En: URL: %5Curl%7B%20https://doi.org/10.1371/journal.pone.0148733%7D.

Selingerová, I., I. Horová y J. Zelinka (2014). "Kernel Estimation of Conditional Hazard Function for Cancer Data". En: *In: Niola V, editor. Recent Advances in Energy, Environment, Biology and Ecology. WSEAS Press*, p. 33-39.

Selingerová, I., S. Katina e I. Horova (2021). "Comparison of parametric and semiparametric survival regression models with kernel estimation". En: *Journal of Statistical Computation and Simulation* 91.13, págs. 2717-2739. DOI: 10.1080/00949655.2021.1906875. URL: https://doi.org/10.1080/00949655.2021.1906875.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman y Hall/CRC. ISBN: 9781138331457. URL: https://plotly-r.com.

Society, American Cancer (s.f.). *Melanoma Skin Cancer*. https://www.cancer.org/es/cancer/cancer-de-piel-tipo-melanoma.html.

Tanner, M. A. y W. H. Wong (1983). "The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Method". En: *The Annals of Statistics* 11.3, págs. 989-993. DOI: 10.1214/aos/1176346265. URL: https://doi.org/10.1214/aos/1176346265.

Therneau, T. M. y P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Venables, W. N. y B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wickham, H. y O'Reilly Media Company Safari (2021). *Mastering Shiny*. O'Reilly Media, Incorporated. ISBN: 9781492047377. URL: https://books.google.com.uy/books?id=ha1CzgEACAAJ.

Xie, Y., J. Cheng y X. Tan (2021). *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.17. URL: https://CRAN.R-project.org/package=DT.

9. Anexo

9.1. Análisis del error cuadrático

Para tener una noción del ancho de ventana óptimo se necesita cierta medida de discrepancia. La forma mas común para medir, en cierto punto x, la *performance* de un estimador por núcleo $\hat{f}(x;H)$ es a través del error cuadrático medio (MSE) , que permite descomponer el error en varianza y sesgo de la siguiente manera:

$$\mathbf{MSE}\left(\hat{f}(x;h)\right) = \mathbf{E}\left(\left[\hat{f}(x;h) - f(x)\right]^2\right) = \mathbf{Var}\left(\hat{f}(x;h)\right) + \mathbf{Sesgo}^2\left(\hat{f}(x;h)\right)$$

donde:

$$\mathbf{Var}\left(\hat{f}(x;h)\right) = \mathbf{E}\left(\hat{f}(x;h)\right)^2 - \left(\mathbf{E}\hat{f}(x;h)\right)^2$$
$$\mathbf{Sesgo}\left(\hat{f}(x;h)\right) = \mathbf{E}\left(\hat{f}(x;h) - f(x)\right)$$

El MSE es una medida de discrepancia local, punto a punto. Cuando estamos estimando una función, el interés generalmente se da en el comportamiento global de $\hat{f}$ como estimador de f. Por lo que en este caso utilizamos una medida de *performance* global, como puede ser la posibilidad de integrar el MSE con respecto a x, para obtener lo que se conoce como Error cuadrático medio integrado (MISE):

$$\mathbf{MISE}\left(\hat{f}(x;h)\right) = \mathbf{E}\int_{R^d}\left(\hat{f}(x;h) - f(x)\right)^2 dx$$

El MISE es la distancia L_2 esperada entre $\hat{f}$ y f. La descomposición de varianza y sesgo del MISE, en varianza integrada (IV) y sesgo cuadrado integrado (ISB), queda de la siguiente forma:

$$\mathbf{MISE}\left(\hat{f}(\cdot;h)\right) = \mathbf{IV}\left(\hat{f}(\cdot;h)\right) + \mathbf{ISB}\left(\hat{f}(\cdot;h)\right)$$

donde:

$$\mathbf{IV}\left(\hat{f}(\cdot;h)\right) = \int_{R^d}\mathbf{Var}\left(\hat{f}(x;h)\right)dx$$
$$\mathbf{ISB}\left(\hat{f}(\cdot;h)\right) = \int_{R^d}\mathbf{Sesgo^2}\left(\hat{f}(x;h)\right)dx$$

El MISE es una cantidad no estocástica que describe la *performance* de un estimador de núcleo, con respecto a una muestra típica de la verdadera densidad.
En alguna situaciones nos interesa más medir el estimador, no en una muestra promedio, sino en los datos que tenemos a mano. En este caso utilizamos una medida de discrepancia estocástica que dependa de los datos, como lo es el error cuadrático integrado, definido como:

$$\mathbf{ISE}\left(\hat{f}(\cdot;h)\right) = \int_{R^d}\left(\hat{f}(x;h) - f(x)\right)^2 dx$$

Mientras que el MISE es un número real, el ISE es una variable aleatoria.