FACULTAD DE MATEMÁTICAS
ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

TRABAJO FIN DE MÁSTER

Modelo de Cox con covariantes dependientes del tiempo

ELENA DE LAS HERAS JIMÉNEZ
MÁSTER EN MATEMÁTICAS

Dirigido por:
D. Juan Manuel Muñoz Pichardo

Sevilla, noviembre de 2023

Índice general

Resumen

El análisis de supervivencia es una metodología de análisis estadístico muy útil en cualquier ámbito de investigación (industria, medicina, economía, biología, demografía. . .) y tiene como objetivo analizar o modelizar el tiempo que tardan en ocurrir uno o más eventos de interés. El modelo más usado en este campo es el modelo semiparamétrico de regresión de Cox, conocido también como modelo de riesgos proporcionales. Una extensión del mismo con grandes aplicaciones prácticas en la ciencia, surge ante la presencia de covariantes dependientes del tiempo. De esta forma, el propósito de este trabajo es el estudio de este último modelo extendido de Cox que incluirá la formulación general para permitir variables dependientes del tiempo, una discusión de las características del modelo y la inferencia estadística sobre el mismo. Además, con objeto de ilustrar la utilidad de este modelo, se usará R y se estudiará dicho modelo sobre datos reales.

Palabras claves: análisis de supervivencia, modelo de regresión de Cox, función de riesgo, hipótesis de riesgos proporcionales, covariante dependiente del tiempo.

Abstract

Survival analysis is a very useful statistical analysis methodology in any field of research (industry, medicine, economics, biology, demography...) and aims to analyze or model the time it takes for one or more events of interest to occur. The most widely used model in this field is the semi-parametric Cox regression model, also known as the proportional hazards model. An extension of this model with great practical applications in science arises in the presence of time-dependent covariates. Thus, the purpose of this paper is the study of the latter extended Cox model which will include the general formulation to allow for time-dependent variables, a discussion of the characteristics of the model and statistical inference on it. In addition, in order to illustrate the usefulness of this model, R will be used and the model will be studied on real data.

Keywords: survival analysis, Cox regression model, hazard function, proportional hazards hypothesis, time-dependent covariates.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En estadística, el análisis de supervivencia se plantea como principal objetivo analizar y modelizar el tiempo transcurrido hasta que se produce un evento o suceso de interés. Dicho análisis tiene utilidad en muchos ámbitos de investigación como pueden ser la medicina, la ingeniería o la sociología. En la medicina, por ejemplo, es habitual el estudio de la variable tiempo hasta que se produce la recaída de un paciente tras suministrarle un tratamiento médico específico para una enfermedad determinada.

En el análisis de supervivencia, la variable objetivo es el tiempo de vida y es imprescindible que se conozca el momento en el que se inicia el estudio de dicha variable en cada caso para así poder realizar un estudio detallado y fiable. Nótese que en algunas ocasiones, puede suceder que los individuos abandonen el estudio antes de que se produzca el evento en cuestión o que se incorporen después del inicio del estudio, de modo que en dichos casos, la información conocida del individuo no es completa, sino parcial. Esta peculiaridad caracteriza al análisis de supervivencia y los datos que la presentan, se conocen como datos censurados.

Existen diferentes técnicas estadísticas para modelar la variable objetivo o de interés, sin embargo, este trabajo se centrará en el estudio del modelo de regresión de Cox o de riesgos proporcionales para predecir el tiempo de fallo o probabilidad de supervivencia a través de diferentes variables que pueden ser influyentes en la variable objetivo y que incluso pueden llegar a depender del tiempo. Además, permite identificar y evaluar la relación que existe entre el conjunto de variables explicativas y la función de riesgo en estudio. Esta técnica multivariante del análisis de supervivencia puede ser útil, por ejemplo, para observar o detectar las variables que influyen a la hora de delinquir de nuevo un preso que ha sido puesto en libertad recientemente (véase Berk et al. (1980)).

Otra técnica usual como la comparación de curvas de supervivencia y la presentación de conceptos básicos del análisis de supervivencia que podrían ser de ayuda para el presente trabajo, se pueden consultar en el Trabajo Fin de Grado *Comparación de curvas de supervivencia con datos censurados* de de las Heras (2021).

1.1. Breve referencia histórica

El origen del análisis de supervivencia se sitúa en el siglo XVII con la creación de tablas de mortalidad que se usaban para realizar estudios demográficos. No obstante, el

concepto de análisis de supervivencia tal y como lo conocemos hoy en día, surgió ligado a la ingeniería con el objetivo de analizar la duración y fiabilidad de los diferentes elementos que componen una máquina. Además, durante la Segunda Guerra Mundial se produjo un gran avance en esta rama estadística para así mejorar la industria militar. Actualmente, este tipo de análisis está muy ligado a la Ciencia de la Salud.

Por su parte, los inicios del análisis de regresión se remontan a finales del siglo XIX en Inglaterra y a las actividades del científico Francis Galton[1] sobre la estatura de los humanos. Para comprender cómo se transmitía la estatura de generación en generación, Galton recopiló datos de la altura de diferentes individuos y de sus padres; y construyó tablas de frecuencia que clasificaban a dichos individuos tanto por su altura como por la altura media de sus padres. De esta forma, el científico llegó a la conclusión de que se podía predecir con cierta exactitud la altura de los individuos en función de la altura de sus progenitores, proponiendo así lo que denominó como «regresión a la media». Estudiando más a fondo su hallazgo, llegó a identificar dos conceptos estadísticos de gran importancia, la regresión y la correlación. El primero de ellos fue con la creación de un método que permitía predecir una variable cuantitativa a partir de los valores de otra variable cuantitativa; y, el segundo concepto con el desarrollo de una función matemática que describiese la relación existente entre las variables involucradas en el modelo. Sin embargo, no es hasta el año 1896 cuando Pearson[2] publica por primera vez un estudio riguroso de la correlación y la regresión en la obra *Philosophical Transactions of the Royal Society of London.*

No obstante, hubo una primera aproximación al análisis de regresión realizado por Boscovich[3] en torno al año 1760 y denominado hoy en día como regresión cuantil, la cual minimiza la suma del valor absoluto de los errores y es la antecedente del método de mínimos cuadrados. Este último método fue publicado por Legendre en 1805 y por Gauss en 1809 tras aplicarlo al problema de la determinación de las órbitas de los cuerpos alrededor del Sol a partir de observaciones astronómicas y se caracteriza por minimizar la suma de los errores al cuadrado.

Concretamente, el modelo de regresión de riesgos proporcionales fue una técnica iniciada por Cox[4] y presentada en 1972 en su artículo *Regression models and life-tables* como una ampliación de los resultados que habían expuesto Kaplan y Meier para la comparación de tablas de mortalidad e incorporación de argumentos de tipo regresivo para su análisis. No obstante, cabe destacar que durante la ponencia en la que presentó los resultados obtenidos de su modelo, resaltó que probablemente, esta nueva técnica tendría más utilidad en los estudios industriales y médicos que en la ciencia actuarial, tal y como ha sucedido en la realidad.

[1] Médico y científico inglés (1822 - 1911) que dedicó sus estudios a la biología, psicología, antropología y estadística.

[2] Historiador, matemático, estadístico y maestro inglés (1857 - 1936) conocido como el padre de la Estadística Aplicada. Introdujo el método de los momentos para la obtención de estimadores y desarrolló la correlación lineal entre otros conceptos y técnicas.

[3] Científico jesuita (1711 - 1787) que destacó en diversas áreas como la matemática, la astronomía, la óptica o la filosofía natural.

[4] Estadístico inglés (1924 - 2022) y miembro de la Royal Statistical Society, conocido por sus contribuciones a la estadística teórica y aplicada, a la bioestadística y a la estadística computacional. Ha sido reconocido con numerosos premios a lo largo de su carrera investigadora y su legado perdura en la investigación estadística actual.

Actualmente, el análisis de supervivencia y el modelo de regresión de Cox, con o sin covariantes dependientes del tiempo, son herramientas muy útiles en muchas áreas de la medicina como la epidemiología, la oncología o la psicología. También son utilizadas en la industria e ingeniería.

1.2. Estructura del trabajo

El presente trabajo se divide en cuatro capítulos con el objetivo de organizar correctamente la información recogida. En este primer capítulo se ha realizado una pequeña introducción al análisis de supervivencia y se han incluido algunas referencias históricas tanto del análisis de supervivencia como de la regresión. Por otro lado, en el Capítulo 2 se introducirá el modelo de riesgos proporcionales o modelo de Cox para sentar las bases de donde parte la técnica que se estudiará en el Capítulo 3. Este modelo plantea como hipótesis estructural una relación entre la función de riesgo y el predictor lineal definido sobre covariantes que se mantienen o pueden considerarse constantes a lo largo del tiempo en el que se desarrolla el estudio. Tal y como se ha indicado, en el Capítulo 3 se describirá una técnica que es una extensión del modelo de Cox que surge cuando hay presencia de covariantes que no se mantienen constantes en el periodo de observación o experimentación del fenómeno bajo estudio, es decir, que "covarían" a lo largo de dicho periodo de tiempo de observación. La descripción contará con la formulación del modelo, la enumeración de sus características y su inferencia estadística. Por último, con el Capítulo 4 se pretende ilustrar la utilidad de la técnica en estudio con datos reales en R. Para ello, se irán presentando los paquetes especializados en el modelo, se describirán las funciones necesarias para el análisis y los resultados obtenidos tras aplicar dichas funciones a los datos en estudio de forma que se ilustre el funcionamiento de estas.

Capítulo 2

Modelo de Cox

El modelo de riesgos proporcionales de Cox, comúnmente conocido como modelo de Cox, es un modelo de regresión que se utiliza en el ámbito del análisis de supervivencia para examinar la relación existente entre la supervivencia de un individuo en estudio y uno o más predictores o covariantes. Tal y como se indica en Palmer (1993), dicho modelo es el más usado para interpretar los efectos que tienen un conjunto de variables predictoras o explicativas sobre la variable tiempo de supervivencia T (variable continua y no negativa) o sobre la función de riesgo $h(t)$ (probabilidad condicional de cambio), ya que las estimaciones obtenidas por este modelo son robustas, consistentes y eficientes. Considérese a partir de ahora la terna $(t_i, \underline{X}_i, \delta_i)$ para cada uno de los n individuos en estudio $(i = 1, \ldots, n)$ donde t_i es el valor de la variable tiempo de supervivencia, $\underline{X}_i = (X_{i1}, \ldots, X_{ip})$ el vector de valores para cada sujeto de las p variables predictoras que se utilizarán para el modelo y δ_i la variable de censura, variable binaria que toma el valor 0 si el dato es censurado y 1 en caso de que el dato sea observado.

Para el desarrollo del presente capítulo, se han consultado diversas fuentes bibliográficas tales como: Kleinbaum y Klein (2012), Moore (2016), Palmer (1993), Liu (2012), Cox y Oakes (1984), Tsiatis (2006) o Lemeshow et al. (2008).

2.1. Formulación del modelo

Dado el vector $\underline{X}_i = (X_{i1}, \ldots, X_{ip})$ de variables explicativas, el modelo de Cox modeliza la tasa de riesgo para el individuo i-ésimo en los siguientes términos:

$$\begin{aligned} h_i(t \mid \underline{X}_i) &= \lim_{h \to 0} \left\{ \frac{P(t \leqslant T < t + h \mid T \geqslant t, \underline{X}_i)}{h} \right\} \\ &= h_0(t) \, \exp(\underline{X}_i' \, \underline{\beta}), \end{aligned}$$

donde $h_0(t)$ es la función de riesgo base[5] (no especificada y no negativa) y $\underline{\beta} = (\beta_1, \ldots, \beta_p)$ el vector de parámetros o coeficientes de regresión que miden la intensidad o fuerza de asociación del aumento o disminución de proporcionalidad de la tasa de riesgo en función

[5]El nombre de función de riesgo base se sostiene en la propiedad del modelo de Cox de que cuando todas las variables explicativas $\underline{X}_i = (X_{i1}, \ldots, X_{ip})$ son nulas, la fórmula del modelo se reduce a $h_0(t)$.

de las covariables. De esta forma, la función riesgo en el instante t viene dado por el producto de dos cantidades, una con dependencia temporal y la otra no. La primera de ellas, $h_0(t)$, solo depende del instante t y la segunda, la exponencial del predictor lineal de las covariantes, $\underline{X}_i' \underline{\beta}$, depende solamente de las p variables predictoras $\underline{X}_i$ que en este caso se consideran independientes de la variable tiempo t, i.e., variables cuyo valor para cada individuo en estudio no varía con el tiempo.

Si se considera $\mathbf{X}$ como la matriz formada por los n vectores de p variables predictoras $\underline{X}_i = (X_{i1}, \ldots, X_{ip})$, se concluye que el vector conjunto de las expresiones de las tasas de riesgos de cada uno de los individuos del estudio dadas por el modelo de Cox es:

$$h(t \,|\, \mathbf{X}) = h_0(t)\, \exp(\mathbf{X}' \underline{\beta}).$$

Como consecuencia de la modelización de la tasa de riesgo a través del modelo de Cox, es posible determinar las variables que tienen relación o influencia sobre la función de riesgo o sobre la función de supervivencia, $S(t \,|\, \mathbf{X})$, dado que también se verifica que:

$$\begin{aligned} S(t \,|\, \mathbf{X}) &= \exp\left(- \int_0^t h(u \,|\, \mathbf{X})\, du \right) \\ &= \exp\left(- \exp(\mathbf{X}'\boldsymbol{\beta}) \int_0^t h_0(u)\, du \right) \\ &= \left[\exp\left(- \int_0^t h_0(u)\, du \right) \right]^{\exp(\mathbf{X}'\boldsymbol{\beta})} \\ &= [S_0(t)]^{\exp(\mathbf{X}'\boldsymbol{\beta})}, \end{aligned}$$

donde $S_0(t)$ es la denominada función de supervivencia base.

2.2. Características del modelo

El modelo de regresión de Cox se caracteriza por ser un modelo semiparamétrico y "robusto" (tal y como se especifica en Kleinbaum y Klein (2012) para denotar que la estimación de los parámetros del modelo son cercanos a sus verdaderos valores); y por poseer lo que se conoce como razón de riesgos proporcionales. Todas estas características hacen que el modelo de Cox sea un modelo popular en el análisis de supervivencia y uno de los más usados. Además, permite la estimación de las curvas de supervivencia asociadas al estudio realizado.

Se trata de un modelo semiparamétrico dado que es un modelo que involucra una componente de dimensión infinita (componente no paramétrica) y otra de dimensión finita (componente paramétrica). La función de riesgo base $h_0(t)$ es la parte de dimensión infinita del modelo, es desconocida, se asume que es continua y suele considerarse como el parámetro de ruido o "molesto". En contraposición, los coeficientes del modelo de regresión β_j para $j = 1, \ldots, p$ son de dimensión finita y son los parámetros de interés del modelo, los cuales se desean estimar. Asimismo, se considera que es un método de estimación robusto, ya que las estimaciones de los coeficientes de regresión del modelo están muy próximas a los verdaderos valores de dichos parámetros si se conociese el verdadero modelo paramétrico que se ajusta a los datos en estudio.

Por último, es destacable lo que se conoce como propiedad de riesgos proporcionales, característica que le da nombre al modelo en estudio. Esta propiedad procede del hecho de que el cociente de riesgos de dos individuos cualesquiera i y k en un instante de tiempo t depende solo del cambio en las variables explicativas y, por tanto, no depende del instante de tiempo considerado:

$$\frac{h_i(t \mid \underline{X}_i)}{h_k(t \mid \underline{X}_k)} = \frac{h_0(t)\, \exp(\underline{X}_i'\underline{\beta})}{h_0(t)\, \exp(\underline{X}_k'\underline{\beta})} = \frac{\exp(\underline{X}_i'\underline{\beta})}{\exp(\underline{X}_k'\underline{\beta})} = \exp(\underline{\beta}(\underline{X}_i' - \underline{X}_k')).$$

Dicho cociente, proporciona de manera sencilla un indicador que mide el efecto que tiene una covariable determinada sobre la tasa de riesgo y se conoce como *hazard ratio* o razón de riesgos.

2.3. Estimación de parámetros

Dada la existencia de datos incompletos en el análisis de supervivencia, los coeficientes del modelo de Cox ($\underline{\beta} = (\beta_1, \ldots, \beta_p)$) no pueden estimarse por el método de máxima verosimilitud ordinario, ya que la función de riesgo es desconocida. Ante esta situación, Cox (1975) propone un método de estimación al que se conoce como método de verosimilitud parcial (*partial likelihood*). Este método se basa en el producto de las verosimilitudes de todos los individuos o sujetos que mueren o fallan sin tener en cuenta aquellas observaciones censuradas, pues se consideran que no aportan información sobre el tiempo de fallo. Técnicamente, Cox utilizó como enfoque de estimación el algoritmo de máxima verosimilitud para una función de verosimilitud parcial. Al igual que en cualquier estimación de parámetros estadísticos, se puede realizar una estimación puntual o una estimación por intervalos de los parámetros del modelo.

En el caso de que no existan tiempos de fallos coincidentes o empates, se consideran de manera ordenada los tiempos de fallos observados:

$$t_1 < t_2 < \cdots < t_D$$

con covariantes asociadas:

$$\underline{x}_{(1)}, \underline{x}_{(2)}, \ldots, \underline{x}_{(D)},$$

respectivamente. Considérese $R(t_i)$ como el conjunto de todos los individuos muestrales en riesgo justo antes del instante t_i incluyendo al conjunto de casos censurados con censura mayor que t_i, i.e., $R(t_i) = \{k : t_k > t_i\}$. En esta situación, se tiene que la verosimilitud parcial para un individuo i se expresa tal y como se muestra a continuación:

$$L^*_{(i)} = P\left[\begin{array}{l|l} \text{un individuo con } \underline{X} = \underline{x}_{(i)} & \text{ha sobrevivido hasta } t_i \text{ y hay un} \\ \text{falle en } t_i & \text{fallo en } t_i \text{ entre los casos de } R(t_i) \end{array}\right]$$

$$= \frac{P\left[\begin{array}{l|l} \text{un individuo con } \underline{X} = \underline{x}_{(i)} & \text{ha sobrevivido hasta } t_i \\ \text{falle en } t_i & \end{array}\right]}{P[\text{hay un fallo en } t_i \text{ entre los casos de } R(t_i)]} = \frac{h_i(t_i \mid \underline{x}_{(i)}; \underline{\beta})}{\sum_{j \in R(t_i)} h_j(t_i \mid \underline{x}_{(j)}; \underline{\beta})}$$

$$= \frac{h_0(t_i)\exp(\underline{x}'_{(i)}\underline{\beta})}{\sum_{j\in R(t_i)} h_0(t_i)\exp(\underline{x}'_{(j)}\underline{\beta})} = \frac{\exp(\underline{x}'_{(i)}\underline{\beta})}{\sum_{j\in R(t_i)}\exp(\underline{x}'_{(j)}\underline{\beta})}.$$

Como consecuencia de la expresión anterior, se concluye que la función de verosimilitud parcial conjunta de la muestra en estudio es:

$$L^*(\underline{\beta}) = \prod_{i=1}^{D} L^*_{(i)} = \prod_{i=1}^{D} \frac{\exp(\underline{x}'_{(i)}\underline{\beta})}{\sum_{j\in R(t_i)}\exp(\underline{x}'_{(j)}\underline{\beta})}. \tag{2.1}$$

De manera equivalente, la expresión (2.1) puede escribirse en términos de la variable de censura δ_i de la siguiente forma:

$$L^*(\underline{\beta}) = \prod_{i=1}^{n}\left[\frac{\exp(\underline{x}'_{(i)}\underline{\beta})}{\sum_{j\in R(t_i)}\exp(\underline{x}'_{(j)}\underline{\beta})}\right]^{\delta_i}.$$

Para estimar los parámetros β_i mediante máxima verosimilitud parcial, hay que maximizar la expresión (2.1) con respecto a cada parámetro del modelo. Sin embargo, tal y como pasa en el caso de la máxima verosimilitud, se maximizará la log-verosimilitud parcial:

$$l^*(\underline{\beta}) = \ln L^*(\underline{\beta}) = \sum_{i=1}^{D}\left[\underline{x}'_{(i)}\underline{\beta} - \ln\left(\sum_{j\in R(t_i)}\exp(\underline{x}'_{(j)}\underline{\beta})\right)\right]. \tag{2.2}$$

De esta forma, los estimadores de $\underline{\beta}$ se obtienen maximizando la función log-verosimilitud parcial (2.2). Para ello, se toman las primeras y segundas derivadas parciales en (2.2) con respecto a cada parámetro del modelo y se igualan a cero. Las primeras derivadas permitirán obtener los posibles valores de los estimadores, mientras que las segundas derivadas se usarán para estudiar qué valores obtenidos de los parámetros son máximos. El sistema de ecuaciones a resolver es el siguiente:

$$\frac{\partial l^*(\underline{\beta})}{\partial \beta_i} = 0 \qquad \forall i = 1,\ldots,p.$$

A continuación, se resuelve el sistema de ecuaciones obtenido usando métodos iterativos como por ejemplo, el método de Newton-Raphson[6] (algoritmo robusto para la función log-verosimilitud parcial). Partiendo de una estimación inicial $\widehat{\underline{\beta}}^{(0)}$, el algoritmo va iterando hasta su convergencia, es decir, hasta que se obtiene cierta estabilidad entre dos iteraciones sucesivas del algoritmo, i.e., $l^*(\widehat{\underline{\beta}}^{(n+1)}) \approx l^*(\widehat{\underline{\beta}}^{(n)})$.

Nótese que cuando en la muestra en estudio existen tiempos de fallos coincidentes o empates (usualmente suceden cuando en el estudio se consideran tiempos de supervivencia discretos y una muestra aleatoria grande), normalmente se aplica el método de Breslow, el de Efron o el *Exact Partial Likelihood* de Cox, tal y como se especifica en Ramírez et al. (2017). En todos ellos, se parte de la siguiente notación: sea $U_{(i)}$ el conjunto de individuos que fallan en el instante t_i; $m_{(i)}$ la multiplicidad del instante t_i o número de fallos en dicho

[6]Procedimiento algorítmico cuyo objetivo es hallar las raíces de una función o sistema de manera iterativa. En general, se caracteriza por poseer una rápida convergencia.

instante t_i, de modo que se verifica que $card\{U_{(i)}\} = m_{(i)}$; $r_{(i)} = card\{R(t_i)\}$ siendo $R(t_i)$ el conjunto de individuos en riesgo en el instante t_i (incluye a los casos censurados con censura mayor que t_i) y $z_{(i)} = \sum_{j \in U_{(i)}} \underline{x}_{(j)}$.

En base a la notación anterior, los dos primeros modelos citados se usan cuando el tiempo es continuo y su complejidad computacional es menor que la del método exacto. A continuación, se enuncian cada uno de los métodos considerados:

- Método de Breslow:

$$L_B^*(\underline{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(z'_{(i)}\underline{\beta}\right)}{\left[\sum_{j \in R(t_i)} \exp\left(\underline{x}'_{(j)}\underline{\beta}\right)\right]^{m_{(i)}}},$$

  donde el numerador tiene en cuenta las covariantes que fallan en el instante t_i, mientras que el denominador, considera la suma de la función de riesgo parcial del conjunto de riesgo $R(t_i)$ elevada a la multiplicidad de cada uno de los instantes t_i considerados.

- Método de Efron:

$$L_E^*(\underline{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(z'_{(i)}\underline{\beta}\right)}{\prod_{j=1}^{m_{(i)}} \left[\sum_{k \in R(t_i)} \exp\left(\underline{x}'_{(k)}\underline{\beta}\right) - \frac{j-1}{m_{(i)}} \sum_{k \in U_{(i)}} \exp\left(\underline{x}'_{(k)}\underline{\beta}\right)\right]},$$

  aproximación más precisa que los métodos exacto y de Breslow y se usa preferiblemente cuando la muestra en estudio es pequeña.

Por último, cuando los tiempos de supervivencia se observan en tiempo discreto, Cox propone el método *Exact Partial Likelihood* que se define de la siguiente manera:

$$L_C^*(\underline{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(z'_{(i)}\underline{\beta}\right)}{\sum_{\mathbf{u} \in U_{(i)}} \exp\left\{\left(\sum_{j \in \mathbf{u}} \underline{x}_{(j)}\right)' \underline{\beta}\right\}}.$$

Tal y como se puede consultar en este apartado, a la hora de estimar los parámetros, solo se ha tenido en cuenta el orden de los tiempos de fallo y censura. Como consecuencia de ello, Harrell (2001) concluye que el modelo de Cox es menos sensible a los outliers del estudio que cualquier modelo parámetrico de análisis de supervivencia. Esta característica favorece el uso del modelo de Cox en problemas de regresión en el ámbito del análisis de supervivencia. Para estudiar en más detalle estas técnicas, consultar Ramírez et al. (2017).

No obstante, la estimación de los parámetros del modelo, también se pueden realizar mediante métodos de remuestreo como las técnicas Bootstrap[7] o Jackknife[8]. Dichas

[7] Técnica de remuestreo que calcula el estimador de un parámetro de interés usando muestras con reemplazamiento a partir de la muestra original y con el mismo tamaño que esta. Se utiliza cuando la distribución de los datos en estudio es difícil de obtener analíticamente o cuando las hipótesis clásicas del problema no se cumplen.

[8] Técnica de remuestreo cuyo objetivo es el cálculo del estimador de un parámetro de interés usando la media de las estimaciones obtenidas al eliminar, para cada una de ellas, una observación de la muestra original en estudio. Se utiliza para la estimación del sesgo y la varianza de diferentes estimadores.

técnicas poseen la ventaja de que reducen la necesidad de asumir determinados modelos probabilísticos para las observaciones. Sin embargo, tiene la desventaja de que, en general, requieren un número elevado de cálculos.

2.4. Contrastes de hipótesis

Como es habitual en los modelos de regresión cuando se estiman los parámetros, se realizan test sobre los parámetros estimados para estudiar si dichos parámetros pueden considerarse significativos o no. Para ello, considérese que, asintóticamente, los parámetros estimados $\underline{\widehat{\beta}}$ siguen una distribución normal de media β y matriz de varianzas y covarianzas $\mathbf{\Sigma} = \Phi^{-1}(\underline{\beta})$ de dimensión $p \times p$, i.e., $\underline{\widehat{\beta}} \overset{a}{\sim} \mathcal{N}_p(\underline{\beta}, \Phi^{-1}(\underline{\beta}))$. A su vez, la media y la matriz de varianzas y covarianzas de la distribución normal en cuestión, pueden ser estimadas de la siguiente manera:

$$\underline{\widehat{\beta}} = \underline{\beta}$$

y

$$\widehat{\mathbf{\Sigma}} = \Phi^{-1}(\underline{\widehat{\beta}}),$$

donde:

$$\Phi(\underline{\beta}) = -\frac{\partial^2 l^*(\underline{\beta})}{\partial \underline{\beta} \partial \underline{\beta}'}.$$

El primer contraste de hipótesis que se estudiará es el contraste de hipótesis individual:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \end{cases}$$

para cada $j = 1, \ldots, p$. Para estudiar la significación de la j-ésima covariante del modelo de Cox considerado, se puede utilizar el estadístico de Wald que se presenta a continuación:

$$Z = \frac{\widehat{\beta}_j}{\sqrt{\widehat{Var}(\widehat{\beta}_j)}},$$

donde $\widehat{Var}(\widehat{\beta}_j)$ es el j-ésimo elemento de la diagonal de $\widehat{\mathbf{\Sigma}}$. Bajo hipótesis nula, Z se distribuye asintóticamente según una distribución normal estándar. En consecuencia, el intervalo de confianza para cada uno de los coeficientes $\widehat{\beta}_j$ se obtiene de la siguiente forma:

$$\widehat{\beta}_j \pm z_{\alpha/2} \sqrt{\widehat{Var}(\widehat{\beta}_j)},$$

dado que bajo la hipótesis nula, el estadístico en estudio Z sigue una distribución normal estándar.

Asimismo, es posible estudiar también el contraste de hipótesis conjunto:

$$\begin{cases} H_0 : \underline{\beta} = (\beta_{1_0}, \ldots, \beta_{p_0})' \\ H_1 : \beta_j \neq \beta_{j_0} \text{ para algún } j = 1, \ldots, p. \end{cases}$$

Por su parte, este último contraste, se puede resolver usando tres métodos o test diferentes: el test de razón de verosimilitud, el test de Wald o el test score. Cada uno de ellos tiene unas propiedades distintas y se basan en estadísticos diferentes. A continuación, se enuncian cada uno de los test citados y recogidos en Cox y Oakes (1984):

- El test de Wald: se basa en la media estimada $(\underline{\widehat{\beta}})$ que sigue asintóticamente una distribución normal y el estadístico considerado es:

$$W_{Wald} = (\underline{\widehat{\beta}} - \underline{\beta}_0)'\Phi(\underline{\widehat{\beta}})(\underline{\widehat{\beta}} - \underline{\beta}_0).$$

  Dicho estadístico, bajo la hipótesis nula del contraste, sigue una distribución χ^2 con p grados de libertad ($W_{Wald} \sim \chi^2_p$).

- El test de razón de verosimilud: compara la función de verosimilitud parcial evaluada en la estimación del parámetro $\widehat{\boldsymbol{\beta}}$, i.e., $L^*(\underline{\widehat{\beta}})$, con esa misma función evaluada en el caso en el que la hipótesis nula del contraste es cierta, $L^*(\underline{\beta}_0)$. El estadístico asociado a este test es el siguiente:

$$W_{LR} = 2(\log L^*(\underline{\widehat{\beta}}) - \log L^*(\underline{\beta}_0)).$$

  Bajo hipótesis nula, el estadístico enunciado sigue una distribución χ^2 con p grados de libertad ($W_{LR} \sim \chi^2_p$).

- Test score: usa las derivadas del logaritmo de la función de verosimilitud parcial evaluada en la hipótesis nula del contraste y el estadístico asociado se enuncia de la siguiente forma:

$$W_{score} = \left(\frac{\partial l^*(\underline{\beta}_0)}{\partial \underline{\beta}}\right)' \left(-\frac{\partial l^{*^2}(\underline{\beta}_0)}{\partial \underline{\beta} \partial \underline{\beta}'}\right)^{-1} \frac{\partial l^*(\underline{\beta}_0)}{\partial \underline{\beta}}.$$

  Bajo hipótesis nula, el estadístico enunciado sigue una distribución χ^2 con p grados de libertad ($W_{score} \sim \chi^2_p$).

Tanto el test de razón de verosimilitud como el test score son contrastes invariantes ante diferentes parametrizaciones, sin embargo, el test de Wald tiene una interpretación más directa que el resto de los test en estudio. Además, el test de razón de verosimilitud converge de manera más rápida a una distribución normal mientras que el test score es más rápido computacionalmente cuando el test a contrastar tiene varios parámetros. No obstante, el test que usualmente se utiliza más es el test de razón de verosimilitud.

2.5. Diagnosis del modelo

La hipótesis fundamental que se debe verificar en el modelo de regresión de Cox es la hipótesis de riesgos proporcionales. Esta hipótesis debe verificarse tanto para cada una de las variables como para el conjunto de variables involucradas en el modelo. En el caso de que no se verifique el contraste para alguna de las variables, se deberá considerar un nuevo modelo de Cox sin dicha variable o realizar alguna transformación para que la variable en cuestión cumpla la hipótesis en estudio. Asimismo, en el caso de que todas las variables cumplan la hipótesis de riesgos proporcionales pero no se cumpla de manera global, será necesario replantear el modelo o realizar alguna transformación para que así se verifique la hipótesis fundamental del modelo.

Tal y como se puede consultar en Kleinbaum y Klein (2012), existen diferentes procedimientos para estudiar el cumplimiento de la hipótesis de riesgos proporcionales: un

procedimiento gráfico, un contraste de bondad de ajuste y un procedimiento que involucra el uso de variables dependientes del tiempo. Se recomienda el uso de al menos dos de los procedimientos que se detallan a continuación a la hora de evaluar si la hipótesis de riesgos proporcionales es válida para el modelo en estudio, pues cada una de las técnicas propuestas tienen sus ventajas e inconvenientes.

2.5.1. Procedimiento gráfico

Como procedimientos gráficos, se encuentran dos posibles técnicas que permiten evaluar si la hipótesis de riesgos proporcionales se cumple o no en el modelo. La primera de ellas consiste en la comparación de las curvas log-log de supervivencia y la segunda, en la comparación de las curvas de supervivencia observadas frente a las esperadas.

- Curvas log-log de supervivencia: se trata de una transformación logarítmica de la curva de supervivencia estimada, en concreto, consiste en tomar dos veces el logaritmo natural de la curva en cuestión. Matemáticamente, la curva log-log de supervivencia se escribe como $-\ln(-\ln \widehat{S})$ y su signo es desconocido. Normalmente, este tipo de gráficos utilizan como estimación de la curva de supervivencia la estimación de Kaplan-Meier. Dado que esta estimación de la curva de supervivencia es escalonada, la curva log-log también lo es. Véase de manera genérica, la construcción de este tipo de curvas: sea la función de supervivencia $S(t \,|\, \mathbf{X})$ tal y como se define a continuación:

$$S(t \,|\, \mathbf{X}) = [S_0(t)]^{\exp(\mathbf{X}'\boldsymbol{\beta})} = [S_0(t)]^{\exp\left(\sum_{j=1}^{p} \beta_j X_j\right)}.$$

  Al tomar un logaritmo natural en dicha función, se obtiene que:

$$\ln\, S(t \,|\, \mathbf{X}) = \exp\left(\sum_{j=1}^{p} \beta_j X_j\right) \cdot \ln\, S_0(t).$$

  Como es bien sabido, se cumple que $0 \leqslant S(t \,|\, \mathbf{X}) \leqslant 1$, por lo que al tomar un segundo logaritmo, se tendrá que:

$$\begin{aligned} \ln[-\ln\, S(t \,|\, \mathbf{X})] &= \ln\left[-\exp\left(\sum_{j=1}^{p} \beta_j X_j\right) \cdot \ln\, S_0(t) \right] \\ &= \ln\left[\exp\left(\sum_{j=1}^{p} \beta_j X_j\right) \right] + \ln[-\ln\, S_0(t)] \\ &= \sum_{j=1}^{p} \beta_j X_j + \ln[-\ln\, S_0(t)], \end{aligned} \tag{2.3}$$

  de modo que el rango de la curva $-\ln(-\ln)$ está comprendido entre $-\infty$ y $+\infty$. Tal y como se puede consultar en la expresión (2.3), la curva log-log de supervivencia se puede expresar como la suma de dos términos, por un lado, un predictor lineal y por otro, la transformación log-log de la función de supervivencia base.

  En el caso de que un modelo de Cox sea adecuado para un conjunto de predictores en estudio, las gráficas empíricas esperadas de las curvas log-log de supervivencia para diferentes individuos de la muestra, deben ser aproximadamente paralelas.

- Curvas de supervivencia observadas frente a las esperadas: se basa en la estratificación de los datos por categorías del predictor a evaluar y posteriormente, la estimación de la curva de supervivencia asociada a cada una de las categorías consideradas. Si para cada categoría del predictor evaluado se tiene que los gráficos observados y esperados están próximos entre sí y no discrepan mucho, entonces se puede concluir que la hipótesis de riesgos proporcionales se cumple. En contraposición, si los gráficos de una o más categorías discrepan entre sí, se llegará a la conclusión de que se está incumpliendo la hipótesis en estudio.

2.5.2. Contraste de bondad de ajuste (GOF)

Se trata de una opción más objetiva que el procedimiento gráfico, dado que con el contraste que se realiza para evaluar si la hipótesis de riesgos proporcionales es cierta o no, se obtiene un estadístico de prueba y un p-valor. En esta sección, se presenta el test de Harrell y Lee (1986) que se trata de una variación de la prueba propuesta originalmente por Schoenfeld (1982) y la cual se basa en los residuos denominados de Schoenfeld asociados a cada uno de los predictores involucrados en el modelo de Cox en estudio y a cada uno de los sujetos que tiene un evento. Dichos residuos se definen como los valores observados de las covariantes o variables explicativas menos los valores esperados en cada instante de fallo y para cada uno de los casos no censurados. Además, se verifica que este tipo de residuos para cada una de las variables predictoras suma cero y que un valor positivo de dicho residuo significa que el valor de la variable asociada es más alto de lo esperado en ese instante de fallo.

A continuación, se enuncia el cálculo de los residuos de Schoenfeld para cada variable predictora o explicativa y caso no censurado:

$$r_{ij}^{(s)} = x_{ij} - \frac{\sum_{k \in R(t_i)} \exp(\underline{x}_k' \widehat{\underline{\beta}}) x_{kj}}{\sum_{k \in R(t_i)} \exp(\underline{x}_k' \widehat{\underline{\beta}})} = x_{ij} - \widehat{\bar{x}}_{(w_i)j},$$

siendo $R(t_i)$ el conjunto de todos los individuos o sujetos muestrales que se encuentran en riesgo justo antes del instante t_i, tal y como se había especificado al comienzo del Apartado 2.3. Se obtiene así una serie de residuos para cada una de las p variables predictoras ($j = 1, \ldots, p$). Si estos residuos mantienen un patrón aleatorio o no sistemático, se puede concluir que existen evidencias de que el efecto de la covariable no cambia respecto del tiempo, hecho que presupone el modelo de Cox. En contraposición, si existiese algún tipo de patrón sistemático en los residuos, se concluirá que el efecto de la covariable cambia a lo largo del tiempo.

De esta forma, se considerará que la hipótesis de riesgos proporcionales es cierta cuando los residuos no muestran tendencias temporales y en la gráfica de los residuos frente al tiempo la pendiente sea nula.

El test en cuestión puede realizarse en tres pasos y estudia si el efecto de cada una de las variables es constante en el tiempo:

1. Creación o ejecución del modelo de Cox y obtención de los residuos de Schoenfeld asociados a cada uno de los predictores, i.e., para cada $j = 1, \ldots, p$, se consideran $\{r_{ij}^{(s)}\}_i$ con $i = 1, \ldots, n$.

2. Creación de una variable que es capaz de clasificar en orden los distintos sujetos en función de los tiempos de fallo: $U(i)$ = rango de t_i en la colección de tiempos de fallo.

3. Estudio de la correlación entre las variables creadas en los dos primeros pasos y se considera como hipótesis nula que la correlación entre los residuos de Schoenfeld y el tiempo de fallo censurado es nulo ($H_0 : \rho_k = 0$). En el caso de que se rechace la hipótesis nula, se concluirá también que la hipótesis de riesgos proporcionales no se cumple.

Dicho test se realiza para cada una de las variables por separado, así como para el conjunto de variables involucradas en el modelo de Cox creado. Además, el estadístico de prueba asociado a este contraste de hipótesis se define de la siguiente forma:

$$Z = \frac{\rho_k \sqrt{n_u - 2}}{\sqrt{1 - \rho_k^2}},$$

donde ρ_k es la correlación entre los residuos de Schoenfeld y el orden de los tiempos de fallo; y n_u es el número total de observaciones no censuradas en la muestra estudiada. Dicho estadístico, bajo la hipótesis nula del contraste, sigue una distribución χ^2 con un grado de libertad cuando se estudia la hipótesis de riesgos proporcionales de manera individual para cada una de las variables involucradas en el modelo y s grados de libertad cuando el estudio de la hipótesis de riesgos proporcionales se hace de manera conjunta sobre todas las variables consideradas en el modelo de Cox creado, siendo s el número de variables involucradas en el modelo bajo estudio.

Nótese que el p-valor obtenido tras realizar este test de bondad de ajuste puede verse influenciado por el tamaño de la muestra con la que se esté trabajando. De esta forma, si la muestra es pequeña, una violación grave de la hipótesis nula puede no ser estadísticamente significativa. Por el contrario, una violación leve de la hipótesis nula puede ser muy significativa. A pesar de este inconveniente, el uso de una prueba estadística ofrece un enfoque más objetivo para evaluar la hipótesis de riesgos proporcionales que el uso de un enfoque gráfico. No obstante, en ocasiones puede ser de utilidad usar un enfoque gráfico dado que permite detectar tipos específicos de desviaciones de la hipótesis de riesgos proporcionales. Se recomienda así el uso tanto de una prueba estadística como de un procedimiento gráfico para concluir de manera definitiva una decisión sobre la violación o no de la hipótesis de riesgos proporcionales.

2.5.3. Uso de variables dependientes del tiempo

En ocasiones, puede ser útil el uso de variables dependientes del tiempo para evaluar la hipótesis de riesgos proporcionales de un modelo de Cox en el que las variables predictoras son independientes del tiempo. Cuando se hace uso de este procedimiento, es necesario considerar el modelo de Cox ampliado en el que se tienen en cuenta las interacciones entre las variables predictoras en estudio y el tiempo, es decir, un modelo de Cox en el que haya involucrados productos de las variables explicativas con alguna función del tiempo (por ejemplo, t, $\log t \ldots$).

Una vez considerado el modelo de Cox ampliado, el estudio de si la hipótesis de riesgos proporcionales es cierta o no para una variable predictora j concreta, se basa en la evaluación de la significación del coeficiente asociado al término producto del modelo de la variable en estudio. En este caso, el modelo se definirá de la siguiente forma:

$$h(t \mid \mathbf{X}) = h_0(t) \exp(\boldsymbol{\beta}\mathbf{X} + \gamma(X_j \cdot g(t))),$$

donde $g(t)$ es la función del tiempo asociada a la variable predictora X_j en estudio. Para concluir si la hipótesis de riesgos proporcionales es cierta para la variable predictora en cuestión, será necesario estudiar el contraste cuya hipótesis nula es $H_0 : \gamma = 0$. En el caso de que el coeficiente sea significativo ($\gamma \neq 0$), se concluirá que la hipótesis de riesgos proporcionales para la variable en estudio no se cumple. Nótese que para el estudio de la significación del coeficiente γ se puede usar el estadístico de Wald o uno de razón de verosimilitud. Ambos, bajo la hipótesis nula siguen una distribución χ^2 con un grado de libertad.

No obstante, también es posible estudiar la validez de la hipótesis de riesgos proporcionales de manera conjunta para varios predictores. Para ello, es necesario estudiar de manera simultánea si todos los coeficientes γ_j del modelo que se presenta a continuación, son todos nulos:

$$h(t \mid \mathbf{X}) = h_0(t) \exp\left(\sum_{j=1}^{p} [\beta_j X_j + \gamma_j (X_j \cdot g_j(t))] \right),$$

donde $g_j(t)$ es la función del tiempo asociada a cada uno de p predictores ($j = 1, \ldots, p$). Debe tenerse en cuenta que diferentes variables predictoras pueden llevar asociadas distintas funciones del tiempo y que la hipótesis nula del contraste a estudiar es $H_0 : \gamma_1 = \cdots = \gamma_p = 0$. Este contraste requiere el uso de un estadístico de razón de verosimilitud que bajo hipótesis nula sigue una distribución χ^2 con p grados de libertad. Además, en caso de que no haya evidencias significativas en contra de la hipótesis nula, se concluirá que la hipótesis de riesgos proporcionales conjunta del modelo de Cox no se cumple. Para más detalles de esta aproximación al problema de la diagnosis del modelo, véase Kleinbaum y Klein (2012).

2.6. Evaluación del modelo

Una vez comprobadas y verificadas las hipótesis de riesgos proporcionales de los predictores involucrados en el modelo, tanto individualmente como conjuntamente, es recomendable evaluar la calidad del modelo creado. Para ello, al igual que sucede en cualquier modelo de regresión, se puede utilizar una medida análoga al coeficiente de determinación R^2 que mida el rendimiento del modelo como la proporción de varianza explicada por dicho modelo. Según Schemper y Stare (1996), en un modelo de riesgos proporcionales no existe una medida única, sencilla, útil y fácil de interpretar y calcular como coeficiente de determinación, dado que tal y como señalan Schemper (1990), Korn y Simon (1990), el coeficiente de determinación R^2 usual es muy sensible a la distribución de los tiempos censurados. Por ello, se presentan a continuación diferentes posibles medidas que actúan como el R^2 de un modelo de regresión lineal.

En primer lugar, se adjunta el coeficiente sugerido por Nagelkerke (1991):

$$R_p^2 = 1 - \exp\left[\frac{2}{n}(l_0^* - l_p^*)\right], \tag{2.4}$$

donde l_p^* es la log-verosimilitud parcial para el modelo ajustado con todas las covariantes (p covariantes), l_0^* es la log-verosimilitud parcial para el modelo sin covariantes y n el número de sujetos o individuos existentes en el estudio. Dicho coeficiente tiene las siguientes propiedades:

- Es consistente con la definición del R^2 clásico de una regresión lineal.
- Es consistente con el método de estimación de máxima verosimilitud.
- Es asintóticamente independiente del tamaño de la muestra, n.
- Puede ser interpretado como la proporción de variación explicada por el modelo.
- Es adimensional, es decir, no depende de las unidades utilizadas.

Otras medidas posibles, son las propuestas por O'Quigley et al. (2005) y Royston (2006). La sugerida por O'Quigley et al. (2005) reemplaza en (2.4) el número de individuos o sujetos n por el número de eventos observados m, de modo que, esta nueva medida depende menos del porcentaje de observaciones censuradas y se conoce como medida de aleatoriedad explicada:

$$R_{p,e}^2 = 1 - \exp\left[\frac{2}{m}(l_0^* - l_p^*)\right].$$

Por último, la medida propuesta por Royston (2006) dependiente de la anterior, tiene como objetivo parecerse más al coeficiente de determinación R^2 para la regresión lineal y sugiere la siguiente formulación:

$$R_{p,v}^2 = \frac{R_{p,e}^2}{R_{p,e}^2 + \frac{\pi^2}{6}(1 - R_{p,e}^2)}.$$

2.7. Variantes del modelo

Además del modelo de Cox estudiado, existen variaciones o modificaciones del mismo para la adaptación de dicho modelo a diferentes situaciones. Entre las modificaciones, destacan el modelo de Cox estratificado, los modelos de fragilidad y el modelo de Cox con covariantes dependientes del tiempo. A continuación, se incluye una breve introducción a cada una de las variantes citadas y posteriormente, en el capítulo siguiente, se profundizará en el modelo de Cox con covariantes dependientes del tiempo, tema principal del presente trabajo:

a) Modelo de Cox estratificado: tal y como se ha presentado, se trata de una modificación del modelo de Cox en el que se controla con el uso de una estratificación de un predictor o predictores, el hecho de que dicho predictor o dichos predictores no verifiquen la hipótesis de riesgos proporcionales. Este tipo de modelo puede ser aplicado para la estratificación de varias variables explicativas en varios estratos o para la estratificación de un único predictor. Partiendo del hecho de que en un modelo de Cox existan r variables $(Z_1, \ldots, Z_r)$ que no satisfacen la hipótesis de riesgos proporcionales y s variables $(X_1, \ldots, X_s)$ que sí las satisfacen de las p variables en estudio, se considerará una nueva variable Z^* a partir de las Z's que se utilizarán para la estratificación. Para cada Z_i se crearán categorías y las combinaciones de dichas categorías conformarán los diferentes estratos correspondientes a las categorías de la variable creada Z^*. De esta forma, para cada estrato, existirá un modelo general de la forma:

$$h_g(t \mid \mathbf{X}) = h_{0g}(t) \exp[\underline{\beta}'\mathbf{X}],$$

donde $g = 1, \ldots, k^*$, con k^* el número de estratos de la variable Z^*, $\underline{\beta} = (\beta_1, \ldots, \beta_s)$ el vector de parámetros del modelo, $\underline{X} = (X_1, \ldots, X_s)$ el vector de variables que cumplen la hipótesis de riesgos proporcionales y h_{0g} la función de riesgo base para cada uno de los estratos considerados. En este caso, se obtienen tantas curvas de supervivencia como estratos tiene la variable Z^* y la forma de estimar los parámetros β's del modelo será a través de la maximización de una función de verosimilitud parcial resultado del producto de todas las k^* funciones de verosimilitud parcial de cada uno de los estratos. Adicionalmente, en este tipo de modelo, se debe estudiar si se cumple o no lo que se conoce como la hipótesis de la no interacción, supuesto que se refiere al hecho de que los s parámetros β del modelo $(\underline{\beta} = (\beta_1, \ldots, \beta_s))$ no varían en los diferentes estratos.

b) Modelos de fragilidad: el modelo de Cox asume que la población en estudio es homogénea, o en otras palabras, que todos los individuos poseen el mismo riesgo de sufrir el evento de interés o independencia de los tiempos de supervivencia de los diferentes sujetos. No obstante, dicha hipótesis no es válida en muchos estudios de forma que es necesario introducir la heterogeneidad no observable (efecto aleatorio) con el uso de una componente de fragilidad. De esta forma, esta componente es capaz de dividir la variabilidad total del estudio en una parte que depende de las variables predictoras, y por tanto, que se puede predecir y otra teóricamente impredecible que es explicada por la componente de fragilidad. En concreto, la fragilidad es una medida de riesgo relativo, ya que cuanto mayor sea la fragilidad de un individuo con respecto al evento o suceso de interés, mayor probabilidad habrá de que se produzca ese evento en dicho sujeto. Este tipo de método es un modelo multiplicativo de riesgo compuesto por tres elementos: la componente de fragilidad, la función de riesgo base y la componente que tiene en cuenta la influencia de las covariables usadas. La componente de fragilidad, representada por Z, asume una distribución de probabilidad (normalmente es considerada la distribución Gamma, dado que ayuda a la hora de maximizar la función de verosimilitud parcial necesaria para obtener los estimadores de los parámetros del modelo) pues se trata de una variable aleatoria y la formulación general del modelo queda de la siguiente manera:

$$h_i(t \mid \underline{X}_i, Z) = Z h_0(t) \exp(\underline{\beta}'\underline{X}_i).$$

c) Modelo de Cox con covariantes dependientes del tiempo: ampliación del modelo de Cox en el que algunas o todas las variables explicativas o predictoras dependen del tiempo. En este caso, la formulación del modelo es:

$$h_i(t \mid \underline{X}_i(t)) = h_0(t) \exp\left[\sum_{j=1}^{p_1} \beta_j X_{ij} + \sum_{k=1}^{p_2} \gamma_k X_{ik}(t)\right],$$

que difiere con la formulación del modelo de Cox sin covariantes dependientes del tiempo en el hecho de que en la parte exponencial, se incluyen tanto los predictores independientes del tiempo, $\underline{X}_i = (X_{i1}, \ldots, X_{ip_1})$, como los predictores dependientes del tiempo, $\underline{X}_i(t) = (X_{i1}(t), \ldots, X_{ip_2}(t))$. Al igual que en el modelo de Cox, los parámetros del modelo se estiman mediante un procedimiento de máxima verosimilitud aunque en este tipo de modelo, los cálculos son más complicados dado que los conjuntos de riesgo no son triviales. Además, en este tipo de modelos, no suele cumplirse la hipótesis de riesgos proporcionales y el riesgo en el instante t de que un individuo sufra el evento de interés, depende del valor de la variable $\underline{X}_i(t)$ en ese mismo momento. En el siguiente capítulo, se profundizará en este modelo, dado que es el tema principal del presente Trabajo Fin de Máster.

2.8. Técnicas de selección de variables

Tal y como indican Guyon y Elisseeff (2003), la selección de variables en un modelo de regresión posee diferentes ventajas en el modelo como la mejora del rendimiento de la predicción de las variables explicativas, la obtención de predictores más rápidos, la reducción de la dimensionalidad de un problema o el evitar la multicolinealidad. En general, el aumento del número de variables predictoras consideradas en un modelo de regresión, lleva consigo un mayor número de parámetros a estimar y una disminución de la precisión individual de cada una de las estimaciones que se traduce en una mayor varianza y en un sobreajuste de la función de regresión estimada. En contraposición, si se consideran menos variables de las que debiera en el modelo, los sesgos aumentan mientras que las varianzas disminuyen, obteniéndose así una mala descripción de los datos. Otro aspecto a considerar, es la correlación existente entre las variables explicativas: si se consideran predictores muy correlacionados entre ellos, la confianza de la predicción del modelo se ve perjudicada. De este modo, los métodos de selección de variables ayudan a encontrar un modelo apropiado que se ajuste bien a los datos y que a su vez, posea un equilibrio entre su bondad de ajuste y simplicidad.

En el ámbito del análisis de supervivencia, se pueden aplicar las mismas técnicas de selección de variables que se usan en cualquier otro modelo de regresión: selección paso a paso, selección de los mejores subconjuntos de covariantes, método *backward*... En el capítulo 5 del libro de Lemeshow et al. (2008) se detallan tres métodos útiles para la selección de variables entre los que se encuentran la selección paso a paso o la selección de los mejores subconjuntos de predictores.

Capítulo 3

Modelo de Cox con covariantes dependientes del tiempo

En el Capítulo 2 del presente trabajo, se ha presentado el modelo de Cox o de riesgos proporcionales en el que solo intervenían predictores, covariantes o variables explicativas independientes del tiempo. Se asumía que las covariables se medían al inicio del estudio y sus valores permanecían fijos a lo largo de todo el estudio. No obstante, en muchas investigaciones de análisis de supervivencia hay involucradas variables explicativas que van evolucionando a lo largo del tiempo. Este tipo de covariantes se denominan variables dependientes del tiempo y pueden ser tanto de naturaleza discreta como continua. En estos casos, el modelo de regresión estudiado no es completamente válido y es necesario realizar ciertas modificaciones que permitan abordar un estudio adecuado del tiempo de supervivencia de los sujetos o individuos en estudio. Para ello, a lo largo de este capítulo, se presentará la extensión de dicho modelo que contará con su formulación, sus características y estimaciones de sus parámetros. Además, también se complementará el estudio con otras consideraciones como los tipos de variables dependientes del tiempo o algunas de las aplicaciones que pueden tener las variables dependientes del tiempo en diferentes estudios de análisis de supervivencia.

3.1. Tipos de covariantes dependientes del tiempo

Una variable dependiente del tiempo es aquella variable cuyos valores cambian a lo largo del tiempo t en estudio. Dentro de las covariantes dependientes del tiempo, se puede hacer una distinción entre dos tipos: las variables internas o endógenas y las variables externas o exógenas.

Por un lado, las variables internas hacen referencia a una característica particular del individuo que solo puede medirse mientras el evento no se haya producido en dicho individuo y no presente censura. Por ejemplo, dentro de este tipo de variables se encuentra la presión arterial de una persona, el tamaño de un tumor o el índice de masa corporal. Todos estos ejemplos reflejan una condición del sujeto y los valores que toman pueden asociarse con la supervivencia del mismo. Asimismo, tal y como indican Zhang et al. (2018), las covariables internas suelen ser el resultado de un proceso estocástico[9] generado por el

[9]Un proceso estocástico es un conjunto de variables aleatorias $\{X_t\}_{t\in T}$ donde t es un parámetro que

individuo bajo estudio y observado solo mientras este sobrevive y no es censurado.

Por otro lado, las variables externas o exógenas son variables dependientes del tiempo que no necesitan que el evento se haya o no producido sobre el individuo en estudio, de modo que, estas variables son independientes del sujeto. Algunos ejemplos de este tipo de variables son: la edad de una persona, las condiciones climáticas de presión atmosférica o los niveles de algún contaminante. Además, este tipo de variables puede influir en la tasa de fallo a lo largo del tiempo. Dentro de las variables externas, existen dos tipos: las definidas y las auxiliares. La primera de ellas, las definidas, son aquellas variables que a priori son independientes del tiempo pero que al ser multiplicadas por una función del tiempo, se obtienen unas nuevas variables dependientes del tiempo. Este tipo de variables definidas se suele usar cuando se quiere verificar si la hipótesis de riesgos proporcionales es cierta para una o varias variables explicativas no dependientes del tiempo en un modelo de Cox creado. La segunda de ellas, las auxiliares, representan las trayectorias observadas de un proceso estocástico en el que la evolución de dicha variable no se ve influenciada por los casos del suceso en estudio. Un ejemplo de este tipo de variables es el nivel de polución del aire.

3.2. Aplicaciones de las covariantes dependientes del tiempo

El uso de covariantes dependientes del tiempo en el modelo de Cox puede tener diferentes aplicaciones, de las cuales, algunas de ellas se pueden consultar en Cox y Oakes (1984). A continuación, se enuncian diferentes aplicaciones de dichas variables:

- Test para la validación de la hipótesis de riesgos proporcionales de un modelo de Cox con variables predictoras independientes del tiempo: tal y como se especificó en el Apartado 2.5.3 del Capítulo 2, la creación de un modelo de Cox ampliado con la presencia del producto de una o todas las variables explicativas del estudio por alguna función del tiempo, puede ser una opción viable para el estudio de la validez de la hipótesis de riesgos proporcionales para dicha o dichas variables explicativas del modelo creado bajo estudio.

- Medidas realizadas a lo largo del estudio: durante la realización de un estudio de análisis de supervivencia, pueden existir variables que estén relacionadas con la supervivencia del evento de interés y que van cambiando a lo largo del tiempo. Este tipo de variables deben ser consideradas en el estudio y el vector formado por ellas se representa de la siguiente forma: $\underline{Z}_i(t) = (Z_{i1}, \ldots, Z_{ip_2})$. Por ejemplo, en investigaciones industriales en las que se estudia la vida útil de una determinada máquina puede ser posible el estudio del desgaste o los daños producidos durante su funcionamiento.

- Cambio de tratamientos: las covariantes dependientes del tiempo pueden usarse para modelizar el efecto de un sujeto que pasa del grupo de control al grupo de tratamiento o viceversa. Un ejemplo sería la evaluación del tratamiento administrado

pertenece a un conjunto T, llamado espacio paramétrico y donde cada variable X_t toma valores reales en $S \subset \mathbb{R}$, denominado espacio de estados.

por las unidades de cuidados intensivos que se usan en sujetos que acaban de sufrir un infarto grave donde el evento de interés es la muerte del sujeto. El riesgo de muerte entre el infarto inicial y el ingreso en la unidad debe tenerse en cuenta en cualquier evaluación del efecto de la unidad de cuidados intensivos sobre la supervivencia. Para ello, se crea una covariable dependiente del tiempo que tomará el valor 0 mientras el paciente no se encuentra en la unidad de cuidados intensivos y 1 en caso contrario.

3.3. Formulación del modelo

Tal y como se indicó en el Apartado 2.5.3 del Capítulo 2, la extensión del modelo de Cox en la que se pueden incluir covariantes dependientes del tiempo modela la tasa de riesgo para el i-ésimo individuo en los siguientes términos:

$$\begin{aligned} h_i(t \mid \underline{X}_i, \underline{Z}_i(t)) &= h_0(t) \exp\left(\underline{X}_i'\underline{\beta} + \underline{Z}_i'(t)\underline{\gamma}\right) \\ &= h_0(t) \exp\left[\sum_{j=1}^{p_1} \beta_j X_{ij} + \sum_{k=1}^{p_2} \gamma_k Z_{ik}(t)\right], \end{aligned}$$

donde $h_0(t)$ es la función de riesgo base (no especificada y no negativa), $\underline{\beta} = (\beta_1, \ldots, \beta_{p_1})$ el vector de parámetros o coeficientes de regresión asociado a las covariantes involucradas en el estudio que son independientes del tiempo, $\underline{\gamma} = (\gamma_1, \ldots, \gamma_{p_2})$ el vector de parámetros o coeficientes de regresión asociado a las covariantes dependientes del tiempo y, $\underline{X}_i = (X_{i1}, \ldots, X_{ip_1})$ y $\underline{Z}_i(t) = (Z_{i1}(t), \ldots, Z_{ip_2}(t))$ son los vectores de las variables independientes y dependientes del tiempo, respectivamente, asociadas a la observación i-ésima.

Si se considera $\underline{X}$ como la matriz formada por los n vectores de p_1 variables predictoras independientes del tiempo $\underline{X}_i = (X_{i1}, \ldots, X_{ip_1})$ y $\underline{Z}(t)$ como la matriz formada por los n vectores de p_2 variables predictoras dependientes del tiempo $\underline{Z}_i(t) = (Z_{i1}(t), \ldots, Z_{ip_2}(t))$, se concluye que el vector conjunto de las expresiones de las tasas de riesgos de cada uno de los individuos del estudio dadas por el modelo de Cox extendido es:

$$h(t \mid \underline{X}, \underline{Z}(t)) = h_0(t)\ \exp(\underline{X}'\underline{\beta} + \underline{Z}'(t)\underline{\gamma}).$$

3.4. Características del modelo

El modelo en estudio es válido para cualquiera de los tipos de variables dependientes del tiempo detalladas en el Apartado 3.1 de este capítulo. Sin embargo, es más apropiado para las variables dependientes del tiempo de tipo externo. Además, una hipótesis importante del modelo es que el efecto que tiene una covariante dependiente del tiempo $\underline{Z}_i(t) = (Z_{i1}(t), \ldots, Z_{ip_2}(t))$ sobre la probabilidad de supervivencia en un instante t depende del valor de dicha variable en ese mismo instante t y no en ningún instante anterior o posterior a dicho t. Además, al igual que el modelo de Cox sin covariantes dependientes del tiempo, es un modelo semiparamétrico, dado que sigue teniendo una componente no paramétrica, la función de riesgo base, $h_0(t)$, la cual se asume que es continua; y dos componentes paramétricas, los coeficientes del modelo β_j para $j = 1, \ldots, p_1$ y γ_k para $k = 1, \ldots, p_2$.

Nótese que, aunque las variables dependientes del tiempo involucradas en el modelo cambian a lo largo del tiempo, la estimación de la tasa de riesgo dada por el modelo de Cox extendido solo ofrece un coeficiente por cada variable del modelo $\underline{Z}_i(t) = (Z_{i1}(t), \ldots, Z_{ip_2}(t))$. De este modo, en el instante t, solo hay un valor de la variable $\underline{Z}_i(t)$ que tiene efecto sobre la tasa de riesgo estimada. Este valor es el que se mide en dicho instante t. No obstante, para evitar esta posible problemática, se puede considerar sobre cada variable lo que se conoce con el nombre de efecto de retardo. Este efecto de retardo consiste en modificar cualquier variable dependiente del tiempo en un periodo de tiempo L_k introducido y elegido por el investigador que está realizando el estudio. De esta forma, se introduce lo que se conoce como modelo de Cox ampliado con retardo de tiempo:

$$h_i(t \mid \underline{X}_i, \underline{Z}_i(t)) = h_0(t) \exp\left[\sum_{j=1}^{p_1} \beta_j X_{ij} + \sum_{k=1}^{p_2} \gamma_k Z_{ik}(t - L_k)\right],$$

en el que se ha sustituido la variable $\underline{Z}_i(t)$ por $\underline{Z}_i(t - L_k)$.

En contraposición a lo que sucedía con el modelo de Cox, este modelo de Cox extendido para covariantes dependientes del tiempo no verifica la propiedad de riesgos proporcionales. Para evidenciarlo, adjuntamos la razón de riesgos para un determinado instante de tiempo y dos individuos cualesquiera i y s que poseen variables tanto dependientes como independientes del tiempo:

$$\begin{aligned}\frac{h_i(t \mid \underline{X}_i, \underline{Z}_i(t))}{h_s(t \mid \underline{X}_s, \underline{Z}_s(t))} &= \frac{h_0(t) \exp\left(\underline{X}_i'\underline{\beta} + \underline{Z}_i'(t)\underline{\gamma}\right)}{h_0(t) \exp\left(\underline{X}_s'\underline{\beta} + \underline{Z}_s'(t)\underline{\gamma}\right)} = \frac{\exp\left(\underline{X}_i'\underline{\beta} + \underline{Z}_i'(t)\underline{\gamma}\right)}{\exp\left(\underline{X}_s'\underline{\beta} + \underline{Z}_s'(t)\underline{\gamma}\right)} \\ &= \exp\left(\left(\underline{X}_i' - \underline{X}_s'\right)\underline{\beta} + \left(\underline{Z}_i'(t) - \underline{Z}_s'(t)\right)\underline{\gamma}\right).\end{aligned}$$

Tal y como se puede comprobar en la expresión anterior, se obtiene que el cociente de riesgos de dos individuos cualesquiera depende del tiempo t. Se concluye así que la razón de riesgos es en función del tiempo. Por lo tanto, se deduce así que, en general, el modelo de Cox extendido para covariantes dependientes del tiempo no satisface la hipótesis de riesgos, salvo que $\underline{\gamma} = \underline{0}$.

Es destacable que en el modelo extendido de Cox para variables dependientes del tiempo, los coeficientes del modelo $\underline{\beta}$ y $\underline{\gamma}$ no dependen de t. Además, en la razón de riesgos, cada uno de los coeficientes γ_k's con $k = 1, \ldots, p_2$ que acompañan a las variables $\underline{Z}_i(t)$ dependientes del tiempo, representan el efecto global que produce la variable en cuestión en cualquier instante de tiempo en estudio.

3.5. Estimación de parámetros

Al igual que en el modelo de Cox sin covariantes dependientes del tiempo, los coeficientes de regresión asociados al modelo de Cox extendido se estiman mediante el procedimiento de máxima verosimilitud. Las estimaciones por máxima verosimilitud se obtienen maximizando la función de verosimilitud parcial asociada al problema en estudio. No obstante, los cálculos para este nuevo modelo extendido son más complejos que el modelo de Cox simple, dado que el conjunto de individuos o sujetos en riesgo es más difícil de considerar cuando se tienen variables que dependen del tiempo, lo cual complica el ajuste computacional y la interpretación de los resultados obtenidos.

A continuación, se presenta el procedimiento de estimación de los parámetros del modelo recogido en Thackham y Ma (2020) para datos censurados a la derecha[10]. Sean y_i para $i = 1, \ldots, n$, los tiempos de supervivencia observados para cada individuo, los cuales tienen asociadas la variable de censura δ_i, de modo que, las observaciones del estudio se representan por la dupla (y_i, δ_i). Tal y como se ha especificado anteriormente, el ajuste del modelo de Cox con covariantes dependientes del tiempo se realiza mediante un enfoque de máxima verosimilitud. Se recuerda que la formulación del modelo de Cox en estudio es:

$$h_i(t \mid \underline{x}_i, \underline{z}_i(t)) = h_0(t) \exp\left(\underline{x}_i'\underline{\beta} + \underline{z}_i'(t)\underline{\gamma}\right), \tag{3.1}$$

donde $\underline{\beta}$ y $\underline{\gamma}$ son los vectores de coeficientes de regresión y $h_0(t)$ es la función de riesgo base no paramétrica que se asume que es continua. Además, para que la función de riesgo expresada en (3.1) sea válida, debe cumplirse que $h_0(t) \geqslant 0$, restricción que debe tenerse en cuenta a la hora de realizar las estimaciones del modelo.

En este modelo, las covariantes de cada uno de los sujetos se dividen en covariantes independientes del tiempo y covariantes dependientes del tiempo, de modo que, el vector de p_1 componentes, $\underline{x}_i = (x_{i1}, \ldots, x_{ip_1})'$, es el que se corresponde con las variables independientes del tiempo y el vector de p_2 componentes, $\underline{z}_i(t) = (z_{i1}(t), \ldots, z_{ip_2}(t))$, es el que contiene a las variables dependientes del tiempo.

Para resolver el problema de la estimación de parámetros, la expresión (3.1) se reescribe de la siguiente forma:

$$h_i(t) = h^*_{oi}(t) \exp\left(\underline{x}_i'\underline{\beta}\right),$$

donde $h^*_{oi}(t) = h_0(t) \exp\left(\underline{z}_i'(t)\underline{\gamma}\right)$ y se estimarán simultáneamente $\underline{\beta}$, $\underline{\gamma}$ y $h_0(t)$. Una de las ventajas de la estimación simultánea es el hecho de que la matriz de covarianza asintótica asociada al problema, se puede establecer con cierta facilidad. La estimación de $h_0(t)$ sin restricciones es inviable, ya que dicho parámetro es de dimensión infinita y solo se dispone un número finito, n, de observaciones. Ante esta situación, una estrategia común consiste en simplificar la función de riesgo base $h_0(t)$ a un subespacio de dimensión finita en el que su dimensión crece con el tamaño de la muestra n, pero a un ritmo más lento. Asimismo, se exige que cuando $n \to +\infty$, $h_0(t)$ converja al verdadero valor de $h_0(t)$ (consúltese Wong y Severini (1991)). El subespacio que se considerará tendrá dimensión $m \leqslant n$ y posee funciones base no negativas $\psi_u(t)$ (donde $u = 1, \ldots, m$) tales que:

$$h_0(t) \approx \sum_{u=1}^{m} \theta_u \psi_u(t). \tag{3.2}$$

Existen diferentes funciones bases no negativas para $\psi_u(t)$, pero en el estudio realizado solo se tratarán dos. La primera de ellas, las funciones indicatrices que dan lugar a una función de riesgo base $h_0(t)$ constante a trozos y, la segunda de ellas, los M-splines[11]. De esta forma, la tasa de fallo acumulada asociada a (3.2) vendrá dada por:

$$H_0(t) = \sum_{u=1}^{m} \theta_u \Psi_u(t),$$

[10] Sea T_i el tiempo de supervivencia del i-ésimo individuo o sujeto y C_i el tiempo de censura, entonces, se tiene que el tiempo de supervivencia observado Y_i para datos censurados a la derecha se define como $Y_i = \text{mín}(T_i, C_i)$.

[11] Conjunto de splines base en el que cada M_i con $i = 1, \ldots, n$, está definido de tal manera que es positivo en (t_i, t_{i+k}) y nulo en el resto de intervalos. Además, son splines normalizados, es decir, $\int M_i \, dx = 1$ y son muy útiles en Estadística.

donde $\Psi_u(t) = \int_0^t \psi_u(s)\, ds$ es la función base acumulada. Como consecuencia, esto implica que la tasa de riesgo acumulada para el i-ésimo sujeto se puede expresar de la siguiente forma:

$$H_i(t) = H_{0i}^*(t) \exp(\underline{x}_i'\underline{\beta}),$$

donde $H_{0i}^*(t) = \sum_{u=1}^{m} \theta_u \Psi_{ui}^*(t)$ y

$$\Psi_{ui}^*(t) = \int_0^t \psi_u(s) \exp(\underline{z}_i'(s)\underline{\gamma})\, ds.$$

La integral anterior se simplifica cuando las covariantes dependientes del tiempo son de tipo discreto. Partiendo de dicha hipótesis, para $t \in (t_{id}, t_{i,d+1}]$, se tiene:

$$\Psi_{ui}^*(t) = \sum_{a=1}^{d} [\Psi_u(t_{i,a+1}) - \Psi(t_{ia})] \exp(\underline{z}_{ia}'\underline{\gamma}), \tag{3.3}$$

donde $\underline{z}_{ia}(t) = \underline{z}_{ia} = [z_{ia1}, \ldots, z_{iaq}]'$ con $\underline{z}_{ib}(t) = \underline{z}_{iab} I(t_{ia} \leqslant t \leqslant t_{i,a+1})$, I la función indicatriz, $a = 1, \ldots, n_i$ con n_i el número de intervalos considerados para el objeto i-ésimo y t_{ia} los "puntos de cambio". Sin pérdida de generalidad, se asume que $t_{i1} = 0$ y $t_{i,n_i+1} = y_i$. De esta forma, cada $z_{ib}(t)$ es constante a trozos con n_i trozos sobre $[0, y_i]$.

Finalmente, se tiene que:

$$H_{0i}^*(t) = \sum_{a=1}^{d} [H_0(t_{i,a+1}) - H_0(t_{ia})] \exp(\underline{z}_{ia}'\underline{\gamma}).$$

Nótese que $\Psi_u(t_{i1}) = 0$ y $H_0(t_{i1}) = 0$ para todo i y u.

La función log-verosimilitud a partir de tiempos de supervivencia independientes $y_1, \ldots, y_n$ es:

$$l(\underline{\beta}, \underline{\gamma}, \underline{\theta}) = -\sum_{i=1}^{n} H_{0i}^*(y_i) \exp(\underline{x}_i'\underline{\beta}) + \sum_{i=1}^{n} \delta_i (\log h_0(y_i) + \underline{x}_i'\underline{\beta} + \underline{z}_{i,n_i+1}'\underline{\gamma}). \tag{3.4}$$

Sea $\underline{\theta}$ un vector con m componentes para todo θ_u. Se desea estimar $\underline{\eta} = [\underline{\beta}', \underline{\gamma}', \underline{\theta}']'$ maximizando la función log-verosimilitud sujeta a las restricciones $\underline{\theta}_u \geqslant 0$ para $u = 1, \ldots, m$, ya que $\psi_u(t) \geqslant 0$. Para ello, es necesario considerar todas las primeras y segundas derivadas parciales en (3.4) con respecto a cada parámetro del modelo e igualarlas a cero.

Para dar solución al problema anterior, se consideran las condiciones necesarias de primer orden de Karush-Kuhn-Tucker (KKT) (Karush (1939), Kuhn y Tucker (1951)) para la solución óptima de los parámetros $\underline{\beta}$, $\underline{\gamma}$ y $\underline{\theta}$:

$$\frac{\partial l}{\partial \beta_j} = 0, \tag{3.5}$$

$$\frac{\partial l}{\partial \gamma_b} = 0, \tag{3.6}$$

$$\frac{\partial l}{\partial \theta_u} = 0 \quad \text{si } \theta_u > 0 \quad \text{o} \quad \frac{\partial l}{\partial \theta_u} < 0 \quad \text{si } \theta_u = 0. \tag{3.7}$$

Existen numerosos algoritmos para la resolución de este tipo de problemas de optimización con restricciones como el método de Newton. Al igual que se expuso en el Apartado 2.3 del Capítulo 2, dichos métodos son iterativos, ya que a partir de una estimación inicial de cada parámetro, el algoritmo va iterando hasta su convergencia.

Sin embargo, cuando m es grande, existe un algoritmo de fácil implementación y muy eficiente para la resolución de las ecuaciones (3.5), (3.6) y (3.7). Dicho algoritmo, se presenta a continuación y comienza estimando $\underline{\beta}^{(k)}$, $\underline{\gamma}^{(k)}$ y $\underline{\theta}^{(k)}$. Una vez obtenidas dichas estimaciones, la iteración $k+1$ se calcula siguiendo los siguientes pasos:

(1) Calcular $\underline{\beta}^{(k+1)}$ tal que $l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k)}, \underline{\theta}^{(k)}) \geqslant l(\underline{\beta}^{(k)}, \underline{\gamma}^{(k)}, \underline{\theta}^{(k)})$.

(2) Calcular $\underline{\gamma}^{(k+1)}$ tal que $l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k+1)}, \underline{\theta}^{(k)}) \geqslant l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k)}, \underline{\theta}^{(k)})$.

(3) Calcular $\underline{\theta}^{(k+1)} \geqslant 0$ tal que $l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k+1)}, \underline{\theta}^{(k+1)}) \geqslant l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k+1)}, \underline{\theta}^{(k)})$.

Estas condiciones aseguran que $l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k+1)}, \underline{\theta}^{(k+1)}) \geqslant l(\underline{\beta}^{(k)}, \underline{\gamma}^{(k)}, \underline{\theta}^{(k)})$ al final de la iteración $k+1$, lo cual es un requisito clave para la convergencia del algoritmo en estudio. Los pasos 1 y 2 se resuelven usando el algoritmo de Newton incorporando pasos de búsqueda lineal[12], y el paso 3, se resuelve mediante un algoritmo multiplicativo-iterativo (MI) diseñado para respetar las restricciones no negativas de $\underline{\theta}$. Para más información sobre este último tipo de algoritmo, consultar Chan y Ma (2012) y Ma (2010). El algoritmo presentado se conoce con el nombre de algoritmo Newton-MI.

Para actualizar el valor de $\underline{\beta}$, se emplea una iteración del algoritmo de Newton con búsqueda lineal. Comenzando con $\underline{\beta}^{(k)}$ y usando una búsqueda lineal con tamaño de paso $\omega_1^{(k)} \in (0, 1]$, se tiene:

$$\underline{\beta}^{(k+1)} = \underline{\beta}^{(k)} + \omega_1^{(k)} (\mathbf{X}'\mathbf{A}^{(k)}\mathbf{X})^{-1}\mathbf{X}'(-\mathbf{A}^{(k)}\underline{1}_n + \underline{\delta}),$$

donde $\mathbf{X}$ es la matriz $n \times p_1$ compuesta por todas las covariantes no dependientes del tiempo, $\mathbf{A}$ es una matriz diagonal definida como

$$\mathbf{A} = diag(H_{01}^*(y_1)\exp(\underline{x}_1'\underline{\beta}), \ldots, H_{0n}^*(y_n)\exp(\underline{x}_n'\underline{\beta})),$$

$\underline{1}_n$ es un vector de 1's de tamaño n y $\underline{\delta}$ es un vector de n componentes para los δ_i's. Además, se tiene que $\mathbf{A}^{(k)}$ es $\mathbf{A}$ con $\underline{\beta} = \underline{\beta}^{(k)}$, $\underline{\gamma} = \underline{\gamma}^{(k)}$ y $\underline{\theta} = \underline{\theta}^{(k)}$. La matriz $\mathbf{X}'\mathbf{A}\mathbf{X}$ es el hessiano negativo de $l(\underline{\beta}, \underline{\gamma}, \underline{\theta})$ con respecto a $\underline{\beta}$, y el parámetro de búsqueda lineal $\omega_1^{(k)}$ ayuda a conseguir que $l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k)}, \underline{\theta}^{(k)}) \geqslant l(\underline{\beta}^{(k)}, \underline{\gamma}^{(k)}, \underline{\theta}^{(k)})$.

Análogamente, el método de Newton con búsqueda lineal se aplica para la actualización de $\underline{\gamma}$. Primero, se considera la matriz $\mathbf{Z}(t) = [\underline{Z}_1'(t), \ldots, \underline{Z}_n'(t)]'$ donde $\underline{Z}_i(t) = [z_{i1}, \ldots, z_{i,n_i+1}]'$ y z_{ia} está definido con la Ecuación 3.3. La matriz $\mathbf{Z}(t)$ es la matriz del modelo asociada a las covariantes dependientes del tiempo y su dimensión es $N \times p_2$ siendo $N = \sum_i n_i$. Sea también

$$\mathbf{B} = diag(\exp(\underline{x}_1'\underline{\beta})\mathbf{B}_1, \ldots, \exp(\underline{x}_n'\underline{\beta})\mathbf{B}_n),$$

[12]Enfoque iterativo y básico en los problemas de optimización que se basa en la búsqueda de una dirección de descenso a lo largo de la cual la función objetivo del problema se reducirá.

donde $\mathbf{B}_i$ es una matriz diagonal de tamaño $n_i \times n_i$ cuyos elementos de la diagonal son $[H_0(r_{i,a+1}) - H_0(r_{i,a})] \exp(\underline{z}'_{ia}\underline{\gamma})$ para $a = 0, \ldots, n_i - 1$. Sea $\underline{\zeta}$ un N-vector dado por $\underline{\zeta} = [\zeta_{11}, \ldots, \zeta_{1,n_1-1}, \ldots, \zeta_{n1}, \ldots, \zeta_{n,n_n-1}]'$ donde $\zeta_{ia} = 1$ solo si $a = n_i - 1$ y $\delta_i = 1$, en cualquier otro caso, $\zeta_{ia} = 0$. De esta forma, el algoritmo de Newton con búsqueda lineal actualiza $\underline{\gamma}$ de acuerdo a la ecuación siguiente:

$$\underline{\gamma}^{(k+1)} = \underline{\gamma}^{(k)} + \omega_2^{(k)}(\mathbf{Z}'(t)\mathbf{B}^{(k)}\mathbf{Z}(t))^{-1}\mathbf{Z}'(-\mathbf{B}^{(k)}\underline{1}_N + \underline{\zeta}),$$

donde $\omega_2^{(k)} \in (0, 1]$ se corresponde con el tamaño del paso de búsqueda lineal y $\mathbf{B}^{(k)}$ denota $\mathbf{B}$ pero con $\underline{\beta} = \underline{\beta}^{(k+1)}$, $\underline{\gamma} = \underline{\gamma}^{(k)}$ y $\underline{\theta} = \underline{\theta}^{(k)}$. La matriz $\mathbf{Z}'(t)\mathbf{B}\mathbf{Z}(t)$ es el hessiano negativo de $l(\underline{\beta}, \underline{\gamma}, \underline{\theta})$ con respecto a $\underline{\gamma}$. Nótese que $\omega_2^{(k)}$ se usa para conseguir que se verifique que $l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k+1)}, \underline{\theta}^{(k)}) \geqslant l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k)}, \underline{\theta}^{(k)})$.

Finalmente, para actualizar el parámetro $\underline{\theta}$, se usa un algoritmo de tipo multiplicativo-iterativo (MI) que respeta la restricción no negativa de $\underline{\theta}$ y, además, es fácil de implementar. Sean $\mathbf{C}$ y $\mathbf{C}^*$ matrices $n \times m$ cuyos elementos (i, u) son $\psi_u(y_i)$ y $\Psi^*_{ui}(y_i)$, respectivamente. Sea $\underline{\delta}$ el vector de dimensión n para δ_i y $\underline{f}$ el n-vector para $\exp(\underline{x}_i\underline{\beta})$. El algoritmo MI actualiza $\underline{\theta}$ de acuerdo con la siguiente ecuación:

$$\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} + \omega_3^{(k)}\mathbf{S}^{(k)}(\mathbf{C}'[\mathbf{D}^{(k)}]^{-1}\underline{\delta} - [\mathbf{C}^{*(k)}]'\underline{f}^{(k)}),$$

donde $\mathbf{D}$ y $\mathbf{S}$ son matrices diagonales cuyos elementos son $h_0(y_i)$ y $\theta_u/(\sum_i \Psi^*_{ui}(y_i) \exp(\underline{x}'_i\underline{\beta}) + \varepsilon)$, respectivamente. Además, ε es un pequeño umbral utilizado para evitar que el denominador correspondiente sea nulo y $\omega_3^{(k)} \in (0, 1]$ para garantizar que se verifique que $l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k+1)}, \underline{\theta}^{(k+1)}) \geqslant l(\underline{\beta}^{(k+1)}, \underline{\gamma}^{(k+1)}, \underline{\theta}^{(k)})$.

Todas las búsquedas lineales utilizadas en el algoritmo, pueden calcularse eficientemente con el uso de la regla de Armijo[13] (Luenberger y Ye (1984)). Asimismo, cuando $\mathbf{A}^{1/2}\mathbf{X}$ y $\mathbf{B}^{1/2}\mathbf{Z}(t)$ tienen rangos máximos, las matrices $\mathbf{X}'\mathbf{A}\mathbf{X}$ y $\mathbf{Z}'(t)\mathbf{B}\mathbf{Z}(t)$ son definidas positivas de modo que las actualizaciones de $\underline{\beta}$ y $\underline{\gamma}$ están bien definidas. Por último, destacar que se puede demostrar que si $\underline{\theta}^{(k)}$ es no negativa, entonces $\underline{\theta}^{(k+1)}$ es también no negativa y que bajo ciertas condiciones de regularidad, el algoritmo expuesto converge a la solución satisfaciendo las condiciones de KKT. Para profundizar en estos últimos resultados, consultar Chan y Ma (2012).

3.5.1. Propiedades asintóticas

Los resultados asintóticos son fundamentales para realizar inferencias sin la necesidad de usar métodos computacionales complejos. Sean $(\underline{\beta}_0, \underline{\gamma}_0, h_{00}(t))$ los verdaderos parámetros y $\widehat{\underline{\beta}}$ y $\widehat{\underline{\gamma}}$ las estimaciones por máxima verosimilitud de $\underline{\beta}$ y $\underline{\gamma}$ respectivamente. Además, se toma la notación siguiente:

$$h_n(t) = \sum_{u=1}^{m} \theta_u \psi_u(t) \quad \text{y} \quad \widehat{h}_n(t) = \sum_{u=1}^{m} \widehat{\theta}_u \psi_u(t),$$

[13]Método esencial en la optimización numérica y búsqueda de línea en problemas de optimización no lineal, especialmente en algoritmos iterativos como el descenso de gradiente, ya que determina el tamaño adecuado del paso a lo largo de una dirección de búsqueda, evitando pasos excesivamente grandes y garantizando la convergencia de los algoritmos de optimización hacia un mínimo local en una función objetivo.

donde $\widehat{\theta}_u$ es la estimación de máxima verosimilitud del parámetro θ_u y m aumenta con n. Sea también $\underline{\theta}$ el vector de todos los θ_u y $\widehat{\underline{\theta}}$ el vector de los $\widehat{\theta}_u$.

Según Xu et al. (2018), se pueden obtener resultados con consistencia fuerte para $\widehat{\underline{\beta}}$, $\widehat{\underline{\gamma}}$ y $\widehat{h}_n(t)$ cuando $m \to +\infty$ y $n \to +\infty$ pero $m/n \to 0$. Además, existen resultados que aseguran la normalidad asintótica de las estimaciones de $\widehat{\underline{\beta}}$, $\widehat{\underline{\gamma}}$ y $\widehat{\underline{\theta}}$ para un m fijo. Asimismo, se puede demostrar que la varianza asintótica es exacta cuando se compara con la varianza obtenida a partir del método de Monte Carlo. Estos últimos resultados se pueden consultar y profundizar en Thackham y Ma (2020).

3.6. Contrastes de hipótesis

Al igual que ocurre en el modelo de Cox con covariantes independientes del tiempo, una vez estimados los parámetros $\underline{\beta}$ y $\underline{\gamma}$ del modelo en estudio, es necesario realizar contrastes para estudiar si dichos parámetros pueden considerarse significativos o no. Dadas las propiedades asintóticas anteriores, se puede considerar que tanto $\widehat{\underline{\beta}}$ como $\widehat{\underline{\gamma}}$ siguen una distribución normal de media $\underline{\beta}$ y $\underline{\gamma}$, y matriz de varianzas y covarianzas $\mathbf{\Sigma}_{\underline{\beta}} = \Phi^{-1}(\underline{\beta})$ de dimensión $p_1 \times p_1$ y $\mathbf{\Sigma}_{\underline{\gamma}} = \Phi^{-1}(\underline{\gamma})$ de dimensión $p_2 \times p_2$, respectivamente. Por lo tanto, se concluye que $\widehat{\underline{\beta}} \overset{a}{\sim} \mathcal{N}_{p_1}(\underline{\beta}, \Phi^{-1}(\underline{\beta}))$ y $\widehat{\underline{\gamma}} \overset{a}{\sim} \mathcal{N}_{p_2}(\underline{\gamma}, \Phi^{-1}(\underline{\gamma}))$. Asimismo, la media y la matriz de varianzas y covarianzas de las distribuciones normales anteriores, se pueden estimar de la siguiente forma:

$$(i) \qquad \widehat{\underline{\beta}} = \underline{\beta} \quad \text{y} \quad \mathbf{\Sigma}_{\underline{\beta}} = \Phi^{-1}(\widehat{\underline{\beta}}) \quad \text{con} \quad \Phi(\underline{\beta}) = -\frac{\partial^2 l^*(\underline{\beta})}{\partial \underline{\beta} \partial \underline{\beta}'},$$

$$(ii) \qquad \widehat{\underline{\gamma}} = \underline{\gamma} \quad \text{y} \quad \widehat{\mathbf{\Sigma}}_{\underline{\gamma}} = \Phi^{-1}(\widehat{\underline{\gamma}}) \quad \text{con} \quad \Phi(\underline{\gamma}) = -\frac{\partial^2 l^*(\underline{\gamma})}{\partial \underline{\gamma} \partial \underline{\gamma}'}.$$

En este modelo, habrá que considerar contrastes de hipótesis individuales para cada uno de los parámetros β_j con $j = 1, \ldots, p_1$ y γ_k con $k = 1, \ldots, p_2$. Estos contrastes se expresan en los siguientes términos:

$$(i) \quad \begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases} \qquad (ii) \quad \begin{cases} H_0 : \gamma_k = 0 \\ H_1 : \gamma_k \neq 0 \end{cases}$$

para cada $j = 1, \ldots, p_1$ y $k = 1, \ldots, p_2$, respectivamente. Dichos contrastes se resuelven utilizando el estadístico de Wald tal y como se indica en el Apartado 2.4 del Capítulo 2.

Además, también es posible considerar el contraste de hipótesis conjunto para cada uno de los parámetros $\underline{\beta}$ y $\underline{\gamma}$. Estos contrastes por separado, se pueden resolver usando los tres métodos y las mismas consideraciones citadas en el Apartado 2.4 anteriormente indicado. A continuación, se presentan los dos contrastes de interés:

$$(i) \quad \begin{cases} H_0 : \underline{\beta} = (\beta_{1_0}, \ldots, \beta_{p_{1_0}})' \\ H_1 : \beta_j \neq \beta_{j_0} \text{ para algún } j = 1, \ldots, p_1, \end{cases}$$

$$(ii) \quad \begin{cases} H_0 : \underline{\gamma} = (\gamma_{1_0}, \ldots, \gamma_{p_{2_0}})' \\ H_1 : \gamma \neq \gamma_{j_0} \text{ para algún } j = 1, \ldots, p_2. \end{cases}$$

Asimismo, es posible considerar el siguiente contraste de hipótesis conjunto de $\underline{\beta}$ y $\underline{\gamma}$:

$$\begin{cases} H_0 : (\underline{\beta}, \underline{\gamma}) = (\beta_{1_0}, \ldots, \beta_{p_{1_0}}, \gamma_{1_0}, \ldots, \gamma_{p_{2_0}})' \\ H_1 : \beta_j \neq \beta_{j_0} \ \text{ o } \ \gamma_k \neq \gamma_{k_0} \ \text{ para algún } j = 1, \ldots, p_1 \ \text{ o } \ k = 1, \ldots, p_2. \end{cases}$$

Este último contraste se suele utilizar en los diferentes software de programación para mostrar la significación del modelo en estudio. Puede resolverse usando por ejemplo, el test de razón de verosimilitud de forma que el estadístico asociado a dicho contraste sería:

$$X_{LR} = 2(\log L^*(\hat{\underline{\beta}}, \hat{\underline{\gamma}}) - \log L^*(\underline{\beta}_0, \underline{\gamma}_0)).$$

Bajo hipótesis nula, el estadístico enunciado sigue una distribución χ^2 con $p_1 + p_2$ grados de libertad ($X_{LR} \sim \chi^2_{p_1+p_2}$).

3.7. Evaluación del modelo

Tal y como ocurre en cualquier modelo de regresión, una vez contrastadas la significación o no significación de los parámetros involucrados en el modelo, es conveniente evaluar la capacidad predictiva del modelo creado. Análogamente a lo que sucedía con el modelo de Cox con covariantes independientes del tiempo, una posible medida para evaluar la calidad del modelo es usando un coeficiente de determinación R^2 adaptado al tipo de datos con el que se trabaja en el análisis de supervivencia.

En el modelo de Cox con covariantes dependientes del tiempo, se pueden considerar las mismas medidas de evaluación sugeridas en el Apartado 2.6 adaptadas al modelo objeto de estudio de este capítulo. Entre ellos, se encuentran:

- El coeficiente sugerido por Nagelkerke (1991) definido como:

$$R^2_p = 1 - \exp\left[\frac{2}{n}(l^*_0 - l^*_p)\right],$$

  donde l^*_p es la log-verosimilitud parcial para el modelo ajustado con todas las covariantes ($p = p_1 + p_2$ covariantes), l^*_0 es la log-verosimilitud parcial para el modelo sin covariantes y n el número de sujetos o individuos existentes en el estudio.

- El coeficiente propuesto por O'Quigley et al. (2005) y conocido como medida de aleatoriedad explicada:

$$R^2_{p,e} = 1 - \exp\left[\frac{2}{m}(l^*_0 - l^*_p)\right],$$

  siendo m el número de eventos observados, l^*_p la log-verosimilitud parcial para el modelo ajustado con todas las covariantes y l^*_0 la log-verosimilitud parcial para el modelo sin covariantes y n el número de sujetos o individuos existentes en el estudio.

- El coeficiente sugerido por Royston (2006) depende del anterior y se define como:

$$R^2_{p,v} = \frac{R^2_{p,e}}{R^2_{p,e} + \frac{\pi^2}{6}(1 - R^2_{p,e})}.$$

3.8. Técnicas de selección de variables

Por último, en ocasiones, puede ser interesante considerar el modelo de Cox con un menor número de variables dado que el aumento del número de variables predictoras consideradas en un modelo de regresión, lleva consigo un mayor número de parámetros a estimar y una disminución de la precisión individual de cada una de las estimaciones, lo que se traduce en una mayor varianza y en un sobreajuste de la función de regresión estimada. De esta forma, el considerar un menor número de variables en el modelo podría evitar la multicolinealidad entre las variables explicativas o la obtención de un modelo con una mayor capacidad predictiva. La solución a estas casuísticas se encuentra con el uso de técnicas dedicadas a la selección de variables como ocurre en cualquier tipo de regresión. Entre estas técnicas, tal y como se comentó en el Apartado 2.8, se encuentran por ejemplo, la selección paso a paso o el método *forward*. Tal y como se especificó en el Capítulo 2, en el libro de Lemeshow et al. (2008) se detallan tres métodos útiles para la selección de variables entre los que se encuentran la selección paso a paso o la selección de los mejores subconjuntos de predictores. Asimismo, se puede hacer uso del enfoque del conjunto activo primal-dual para resolver el problema de la selección de subconjuntos de variables y propuesto por Ito y Kunisch (2013). Para profundizar en esta técnica, consultar Jareño (2022).

En general, los métodos de selección de variables ayudan a encontrar un modelo apropiado que se ajuste bien a los datos y que a su vez, posea un equilibrio entre su bondad de ajuste y simplicidad.

Capítulo 4

Modelo de Cox en R

En este capítulo, se detallará cómo se puede usar el modelo de Cox tanto con covariantes independientes como dependientes del tiempo en el software libre R. Para ello, se irán exponiendo las librerías o paquetes necesarios para realizar dicho estudio y las funciones imprescindibles para obtener las conclusiones oportunas. Para mostrar cada uno de los modelos de Cox considerados, se detallarán los argumentos y la salida de cada una de las funciones más relevantes. Asimismo, se adjuntarán ejemplos con datos disponibles en los propios paquetes o con datos extraídos de la asignatura Modelado y Predicción Estadística. Cabe destacar que se usará un nivel de significación del 5 % en todos los contrastes a realizar.

4.1. Introducción

La librería más habitual a la hora de tratar datos de análisis de supervivencia es la librería *survival* (Therneau (2023)). Dicha librería posee funciones esenciales para el análisis de supervivencia y trabaja con objetos de tipo *Surv*. Este tipo de objetos se caracteriza por combinar información temporal y de censura para cada uno de los individuos u observaciones en estudio; y se puede crear con la función *Surv(time, time2, event, type, origin)* cuyos argumentos se enuncian a continuación:

- *time*: en el caso de que los datos estén censurados a la derecha, en este argumento se anotará el tiempo de seguimiento. En contraposición, cuando la censura sea por intervalos, se introducirá el instante inicial del intervalo.
- *event*: indicador del estado de la observación o individuo en estudio, 0 = *vivo* y 1 = *muerto*. No obstante, cuando se trabaja con datos censurados por intervalos, el estado puede ser 0 = *censura a la derecha*, 1 = *evento observado*, 2 = *censura a la izquierda* o 3 = *censura por intervalos*.
- *time2*: argumento necesario cuando se trabaja con datos censurados por intervalos (solo admite intervalos semiabiertos de la forma $(inicio, fin]$) y debe anotar el instante final del intervalo. Además, cuando se trabaja con datos de procesos de conteo el argumento *event* indica si se produjo el evento al final del intervalo.

- *type*: indicador del tipo de censura y admite los siguientes valores *"right"*, *"left"*, *"counting"*, *"interval"*, *"interval2"* o *"mstate"*. Para una mayor explicación de los posibles valores del argumento en estudio, consúltese Therneau (2023).

- *origin*: útil cuando se trabaja con datos de procesos de conteo e indica la función de riesgo al comienzo del estudio.

No obstante, dicha librería no va a ser la única que se va a utilizar a lo largo del presente capítulo. Otras librerías útiles para el estudio del modelo de Cox en R son la librería *survMisc* (Dardis (2022)), la librería *survminer* (Biecek et al. (2021)) o la librería *glmnet* (Friedman et al. (2023)). Por ejemplo, la primera de ellas permite consultar, entre otros conceptos del modelo de Cox, los diferentes coeficientes R^2 expuestos a lo largo del trabajo; y, la segunda, ayuda a presentar gráficos detallados.

4.2. Modelo de Cox

Para ajustar un modelo de Cox en R a partir de ciertas variables explicativas o covariantes, se usa la función *coxph()* de la librería *survival* (Therneau (2023)). Esta función *coxph()* permite crear el modelo de Cox con las covariantes que se le especifique en dicha función a partir de una fórmula. Los argumentos de dicha función se detallan a continuación:

```
library(survival)
args(coxph)
```

```
## function (formula, data, weights, subset, na.action, init, control,
##     ties = c("efron", "breslow", "exact"), singular.ok = TRUE,
##     robust, model = FALSE, x = FALSE, y = TRUE, tt, method = ties,
##     id, cluster, istate, statedata, nocenter = c(-1, 0, 1), ...)
## NULL
```

donde:

- *formula*: objeto de tipo fórmula en el que la parte de la derecha contiene un objeto de tipo *Surv()* seguido por ~ y las variables explicativas o predictoras que se quieran utilizar.

- *data*: conjunto de datos que contiene las variables objeto de estudio para el modelo de Cox.

- *weights*: vector de ponderaciones que se utilizará en el caso que corresponda.

- *subset*: subconjunto de datos que debe ser utilizado en el ajuste a realizar. Es muy útil cuando se trabaja con conjunto de datos de entrenamiento y test.

- *na.action*: función que filtra los datos de tipo *NA*.

- *init*: vector de valores iniciales de la iteración del método que calcula los coeficientes del modelo.
- *control*: objeto de clase *coxph.control* que especifica el límite de iteraciones y otro tipo de opciones de control del ajuste del modelo de Cox.
- *ties*: cadena de caracteres que especifica el método a utilizar en caso de empates en los tiempos de fallo. En el caso de que no existan tiempos de fallo coincidentes, cualquier método es equivalente. Por defecto, se utiliza el método de Efron "*efron*" (más preciso y computacionalmente más eficiente) pero otras opciones posibles son el método de Breslow "*breslow*" o el método *Exact Partial Likelihood* "*exact*", aconsejable cuando se trabaja con un conjunto pequeño de valores discretos.
- *singular.ok*: valor lógico que indica qué hacer en caso de colinealidad en la matriz del modelo. Por defecto, toma el valor *TRUE* de manera que el programa no considera las columnas de la matriz X que son combinaciones lineales de las columnas restantes. En dicho caso, los coeficientes para dichas columnas serán *NA* y la matriz de varianzas contendrá ceros. Además, para los cálculos auxiliares como el predictor lineal, los valores perdidos se tratan como valores nulos.
- *robust*: valor lógico que garantiza el cálculo de una varianza robusta en el caso de tomar el valor *TRUE* (valor por defecto).
- *id*: variable opcional que se utiliza para identificar a los sujetos o individuos de la muestra cuando existen varias filas que hacen referencia al mismo individuo.
- *cluster*: variable opcional que agrupa las observaciones en clústeres y se utiliza cuando el argumento *robust* toma el valor *TRUE*. Normalmente, esta variable se encuentra en los datos de estudio.

Para un estudio más detallado sobre el resto de argumentos que admite la función en estudio, consultar la función *coxph()* y el apartado *coxph.control* del manual de Therneau (2023).

En la salida que proporciona la función se podrá consultar una tabla con la siguiente información para cada una de las variables explicativas involucradas en el modelo:

- *coef*: coeficientes estimados asociados a cada uno de las variables predictoras.
- *exp(coef)*: exponencial de los coeficientes estimados y se interpretan como efectos multiplicativos sobre la función de riesgo $h(t)$.
- *se(coef)*: errores de estimación asociados a los coeficientes estimados.
- z y p: muestran la significación de cada una de las variables consideradas en el modelo creado. El parámetro z muestra el valor del estadístico de Wald asociado a cada variable junto con el p que se corresponde con el p-valor asociado. En función del valor de p se podrán considerar nulos o no los coeficientes de las variables del modelo en estudio.

Asimismo, en la salida, se puede consultar también el estadístico de razón de verosimilitud del contraste de hipótesis conjunto del modelo junto con su grado de libertad correspondiente y p-valor. De esta forma, de manera sencilla se puede concluir si el modelo de Cox considerado es óptimo o se debe ajustar con menos variables explicativas u otras variables diferentes. La salida también adjunta el tamaño de la muesta en estudio, n, el número de eventos que se han producido a lo largo del tiempo de estudio y el número de observaciones que se han eliminado por ser valores perdidos.

No obstante, un estudio más completo de la salida se puede obtener ejecutando la orden *summary()* del modelo creado. En este caso, se muestra una tabla en la que además de la información expuesta anteriormente, se recoge para cada una de las variables explicativas consideradas en el modelo, el valor de *exp(-coef)*, el extremo inferior (*lower .95*) y superior (*upper .95*) del intervalo de confianza asociado a cada una de las exponenciales de los coeficientes estimados cambiados de signo e información relevante de cada uno de los test (valor del estadístico, grados de libertad asociado y p-valor) de Wald, de razón de verosimilitud y test score correspondientes al contraste de hipótesis conjunto del modelo. Por defecto, el intervalo de confianza se calcula al 95 %. En el caso de que se quiera conocer el valor del intervalo de confianza asociado directamente a los coeficientes estimados, se debe usar la función *confint()*. La estructura de dicha función es:

confint(x, parm, level),

donde x es el objeto *coxph* que contiene el modelo de Cox creado, *parm* es el vector que contiene el nombre de las variables para las cuales se quiere estudiar el intervalo de confianza de su coeficiente estimado y *level* indica el nivel de confianza para el cual se quiere consultar el intervalo de confianza (por defecto, toma el valor 0.95). La salida de la función es una matriz (o vector) con dos columnas: en la primera de ellas se muestra el límite inferior del intervalo de confianza y en la segunda, el límite superior de este. Ambas columnas se etiquetan como (1 - nivel)/2 y 1 - (1 - nivel)/2 en porcentaje.

En ocasiones, puede ser de interés observar el comportamiento de la estimación de los coeficientes del modelo de Cox creado, para ello, existe la orden *profLik()* de la librería *survMisc*. Dicha orden presenta como salida una gráfica por cada una de las variables explicativas contempladas en el modelo de Cox bajo estudio en la que se puede consultar la evolución del valor de la función de verosimilitud parcial frente al coeficiente estimado para cada una de las iteraciones realizadas en el ajuste del modelo. Además, en cada gráfica se puede consultar el intervalo de confianza asociado al coeficiente estimado (máximo de la función dibujada) a través de dos circunferencias que denotan el extremo inferior y superior de este. El modo en el que se usa la función en estudio es:

profLik(x, CI, interval, devNew, ...),

siendo x el objeto *coxph* que contiene el modelo de Cox bajo estudio, *CI* el nivel a usar para mostrar el intervalo de confianza (por defecto, toma el valor 0.95), *interval* el número de puntos sobre los que evaluar el coeficiente (por defecto, toma el valor 50), *devNew* el valor lógico para indicar si las gráficas se quieren mostrar en otra pestaña o no y parámetros gráficos adicionales como *col* para indicar el color de la gráfica a dibujar.

Por último, destacar también que la clase de objeto *coxph* contiene más información que se puede consultar para la evaluación del modelo como la matriz de covarianzas de los estimadores, el valor de la log-verosimilitud del modelo con los valores iniciales y finales de los coeficientes, las medias de las variables predictoras, el número de iteraciones

realizadas para la estimación de los coeficientes[14] o el vector de predictores lineales de las observaciones del conjunto de estudio. Para conocer con más detalle la información que contiene el objeto de tipo *coxph*, consultar el apartado *coxph.object* del manual de Therneau (2023). Además, es posible realizar un estudio de los residuos de Schoenfeld asociados a cada observación del conjunto de datos bajo estudio con la orden

$$residuals(object,\ type\ =\ c("schoenfeld")),$$

siendo *object* el modelo de Cox creado.

En el caso de que se quiera realizar un estudio gráfico del modelo de Cox creado, existen funciones como *ggforest()* del paquete *survminer* que pueden ayudar a realizarlo. Esta función tiene la peculiaridad de mostrar gráficamente la exponencial de los coeficientes obtenidos en el modelo ajustado junto con el intervalo de confianza obtenido para este, así como presentar un pequeño resumen que contiene el p-valor asociado a cada uno de los contrastes individuales de los coeficientes del modelo y al contraste conjunto, el AIC del modelo o el número de observaciones asociadas a cada una de las variables predictoras del modelo. La estructura de la función es:

$$ggforest(model,\ data,\ main,\ cpositions,\ fontsize,\ refLabel,\ noDigits)$$

y cuyos argumentos se definen a continuación:

- *model*: objeto de tipo *coxph*.
- *data*: conjunto de datos usado para ajustar el modelo en estudio.
- *main*: título del gráfico a realizar.
- *cpositions*: posición relativa en el eje OX de las tres primeras columnas que se presentan en la gráfica.
- *fontsize*: tamaño de la fuente utilizada para realizar las anotaciones que aparecen en la gráfica.
- *refLabel*: etiqueta de referencia para las variables de tipo factor.
- *noDigits*: número de dígitos para las estimaciones y los p-valores presentados en el estudio realizado.

Una vez creado un modelo de Cox, también se puede hacer un estudio de las curvas de supervivencia ajustadas para el modelo de regresión creado. Para dicho objetivo, existen dos órdenes en el paquete *survminer*: *ggadjustedcurves()* y *surv_adjustedcurves()*. La primera de ellas, realiza una gráfica de la estimación de las curvas de supervivencia del modelo de Cox creado con la función *coxph()*, mientras que la segunda de ellas, la estima y la muestra de manera tabular en cada instante de tiempo t. Ambas órdenes comparten argumentos y estos, se detallan en la siguiente lista:

- *fit*: objeto de tipo *coxph* que contiene el modelo de Cox creado.

[14] El ajuste del modelo de Cox se puede controlar con diferentes parámetros numéricos de la orden *coxph.control()* como el número máximo de iteraciones para la convergencia del modelo.

- *variable*: nombre de la variable que se va a trazar. Este argumento es útil cuando se realiza un modelo de Cox estratificado. Por defecto, traza la curva media de la población total.

- *data*: conjunto de datos para predecir. En el caso de que no se indique conjunto de datos, se tomará el conjunto de datos usado para el ajuste del modelo.

- *method*: argumento que describe qué tipo de curva de supervivencia debe ser estimada. Las posibles opciones son: *"single"* (calcula y representa una única curva de supervivencia que se corresponde con la curva de supervivencia esperada de los datos proporcionados a partir del ajuste del modelo realizado), *"average"* (representación de una curva de supervivencia por cada nivel o estrato de la variable explicativa especificada), *"marginal"* (análogo al anterior) o *"condicional"*. El resto de argumentos posibles (muchos de ellos, referentes a propiedades de formato de las gráficas como el tema o la paleta de colores disponible) y su detalle se puede consultar en el manual Biecek et al. (2021).

4.2.1. Diagnosis del modelo de Cox

Tal y como se ha indicado en la parte teórica, la diagnosis del modelo se basa en el estudio de si la hipótesis de riesgos proporcionales se verifica o no en el modelo de Cox creado. Para ello, en R se utilizarán funciones destinados al estudio de las hipótesis tanto de manera gráfica como analítica mediante el test de bondad de ajuste (GOF).

En primer lugar, se presenta la función *cox.zph()* del paquete *survival* la cual proporciona una salida en la que se pueden consultar los test de bondad de ajuste individuales de cada una de las variables explicativas contempladas en el modelo, así como el test de bondad de ajuste global o conjunto de todas las variables predictoras. De esta forma, de manera sencilla se puede evaluar si las variables predictoras del modelo cumplen la hipótesis de riesgos proporcionales tanto individualmente como conjuntamente. La sintaxis de la función

$$cox.zph(fit,\ transform,\ global,\ \ldots),$$

se detalla a continuación con el significado de cada uno de sus argumentos:

- *fit*: objeto de tipo *coxph* que contiene el modelo de Cox bajo estudio.

- *transform*: cadena de caracteres que especifica el tipo de transformación a usar en los tiempos de supervivencia antes de realizar el test. Los valores posibles son *"km"* (valor por defecto y realiza una transformación de tipo Kaplan-Meier), *"rank"* (transformación de rango), *"identity"* (no aplica ningún tipo de transformación) o cualquier otra función de un solo argumento.

- *global*: valor lógico que indica si se desea obtener el resultado de realizar el contraste de hipótesis de riesgos proporcionales conjunto de todas las variables predictoras o no. Por defecto, toma el valor *TRUE*.

En el caso de que se quiera profundizar en los argumentos de la función *cox.zph()* consultar el apartado de dicha función en el manual Therneau (2023). En la salida principal

de la función en estudio, *cox.zph()$table*, se obtiene una matriz en la que se puede consultar el valor del estadístico χ^2 (*chisq*) asociado a cada uno de los contrastes junto a sus grados de libertad (*df*) y el p-valor (*p*) correspondiente a cada uno de los test realizados. Además, se puede extraer la matriz de varianzas y covarianzas de las variables predictoras en *cox.zph()$var* y la matriz de residuos de Schoenfeld para cada uno de los tiempos observados (filas) y variables explicativas del modelo (columnas) usando *cox.zph()$y*. Asimismo, si se quiere realizar un test gráfico de la hipótesis de riesgos proporcionales, se puede usar la orden

plot(x, resid, se, df, var),

de forma que se obtiene para cada una de las variables explicativas del modelo, una gráfica con los residuos de Schoenfeld frente al tiempo junto con una curva suavizada y ajustada a través de splines. Se dirá que la variable explicativa cumple la hipótesis de riesgos proporcionales cuando la pendiente de la recta dibujada sobre los residuos es nula. Los argumentos de la orden en estudio son:

- *x*: objeto de tipo *coxph* que contiene el modelo de Cox ajustado.
- *resid*: valor lógico que se utiliza para indicar si se quiere incluir en la gráfica tanto los residuos como la curva ajustada. Por defecto, toma el valor *TRUE*.
- *se*: valor lógico utilizado para indicar si se desean incluir las bandas de confianza de la curva ajustada. Por defecto, toma el valor *TRUE*.
- *df*: grados de libertad de los splines naturales ajustados. Si se desea un ajuste lineal, se usará *df = 2*.
- *var*: vector en el que se indica el conjunto de variables para las que se quiere mostrar la gráfica en estudio.

También se puede configurar algunos aspectos de la gráfica como el color, el título del eje OX o el grosor de la curva ajustada. Para más información sobre la orden y sus argumentos, consultar el apartado *plot.cox.zph* del manual de Therneau (2023).

Las gráficas descritas de los residuos de Schoenfeld para cada una de las variables explicativas consideradas en el modelo y realizadas con la orden *plot()* pueden obtenerse también con la función *ggcoxzph()* de la librería *survminer*. En este caso, además de las gráficas de los residuos, se adjunta para cada variable el p-valor obtenido al realizar el test de bondad de ajuste, así como el p-valor asociado al test de bondad de ajuste global. De esta forma, con la misma orden es posible evaluar si es cierta la hipótesis de riesgos proporcionales de dos maneras distintas (gráficamente y analíticamente). La sintaxis de la función

ggcoxzph(x, resid, se, df, var)

es la misma que se ha descrito en *plot.cox.zph* pero con la diferencia de que se añaden argumentos gráficos de la librería *ggplot2* (Wickham et al. (2023a)). En el caso de que se quiera ampliar información sobre la función en cuestión y sus argumentos, consultar el apartado *ggcoxzph* del manual de Biecek et al. (2021).

Por último, el estudio de la hipótesis de riesgos proporcionales se puede realizar de manera gráfica con otra función de la librería *survminer* denominada *ggcoxdiagnostics()*.

Esta función puede dibujar tanto los residuos de Schoenfeld como otros tipos de residuos (los de Martingala). Dado que el resto de residuos no se contemplan en el estudio realizado a lo largo de este trabajo, solo se enunciará la forma en la que se dibujan los residuos de Schoenfeld. La estructura de la función es

$$ggcoxdiagnostics(fit,\ type,\ ...),$$

donde *fit* es un objeto de tipo *coxph* que contiene el modelo de Cox bajo estudio y *type* es el tipo de residuos que se quiere mostrar en la gráfica y que para el presente trabajo, tomará el valor *"schoenfeld"*. Además, se pueden añadir otros argumentos para modificar el aspecto de la gráfica (consúltese el apartado *ggcoxdiagnostics* del manual de Biecek et al. (2021)). Como salida, se obtienen tantas gráficas como variables explicativas se hayan contemplado en el modelo de Cox creado, en las que se puede consultar la evolución de los residuos de Schoenfeld a lo largo del tiempo. En función del patrón que muestren los residuos, se concluirá de manera gráfica si las variables predictoras del modelo cumplen o no la hipótesis de riesgos proporcionales.

Asimismo, se pueden estudiar las curvas log-log de supervivencia para la diagnosis del modelo. En dicho caso, un modelo de Cox será adecuado para un conjunto de predictores en estudio cuando las gráficas empíricas esperadas de las curvas log-log de supervivencia sean aproximadamente paralelas. La forma en la que se consiguen dichas gráficas es mediante el uso de la función

$$ggsurvplot(fit,\ fun\ =\ "cloglog")$$

del paquete *survminer* donde *fit* es un objeto de tipo *survfit()* y *fun = "cloglog"* denota el tipo de gráfica que se quiere obtener.

4.2.2. Evaluación del modelo de Cox

Para evaluar la calidad del modelo de Cox, se puede usar la función *rsq(x)* de la librería *SurvMisc* (Dardis (2022)) donde x es un objeto *coxph* que contiene el modelo de Cox creado. Como resultado, se pueden obtener los tres coeficientes de determinación R^2 estudiados en la parte teórica del modelo expuestos en los dos capítulos anteriores (véanse los apartados 2.6 y 3.7):

- *cod*: muestra el coeficiente sugerido por Nagelkerke y que se ha denotado por R^2_p.
- *mer*: adjunta lo que se conoce como medida de aleatoriedad explicada y que se ha definido como $R^2_{p,e}$.
- *mev*: muestra la medida propuesta por Royston y que se ha denotado por $R^2_{p,v}$.

Por defecto, la orden devuelve los coeficientes redondeados pero dicha salida puede modificarse con el uso del argumento *sigD* seguido del número de decimales que se desean mostrar. Además, en el caso de que se quieran obtener los coeficientes originales calculados, se añadirá $sigD = NULL$.

4.2.3. Selección de variables del modelo de Cox

Para hacer un estudio de selección de variables en un modelo de Cox, se puede usar la librería *glmnet* (Friedman et al. (2023)) y la función del mismo nombre, *glmnet()*. Dicha librería, tiene como objetivo el ajuste de modelos lineales generalizados mediante la técnica de máxima verosimilitud penalizada[15]. En concreto, la librería en estudio utiliza una penalización de tipo *elastic net* para el modelo de Cox, es decir, una combinación de las penalizaciones *lasso* y *ridge*; y resuelve el problema usando el algoritmo de descenso por coordenadas cíclicas[16].

La función *glmnet(x, y, family, nlambda, lambda,...)* permite ajustar el modelo de Cox, así como otros modelos lineales generalizados, a través de la máxima verosimilitud penalizada con el objetivo de obtener un modelo de Cox sencillo y fácil de interpretar. Los argumentos de la función en estudio se detallan a continuación:

- x: matriz que contiene para cada una de las observaciones en estudio (filas) el valor asociado de las variables explicativas contempladas en el modelo de Cox a ajustar.
- y: matriz de dos columnas que contiene para cada una de las observaciones en estudio el valor t junto con el estado de censura de cada una de ellas u objeto de tipo *Surv*.
- *family*: cadena de caracteres que representa la familia a la que pertenece el modelo que se quiere ajustar. Puede tomar los valores *"gaussian"*, *"binomial"*, *"poisson"*, *"multinomial"*, *"cox"* o *"mgaussian"*, siendo el valor *"cox"* el que compete a este estudio.
- *nlambda*: número de valores de *lambda* a utilizar en el ajuste. Por defecto, toma el valor 100.
- *lambda*: secuencia de valores del coeficiente de penalización asociado al ajuste del modelo a construir.

Existen otros muchos más argumentos en la presente función. Para un estudio más detallado de ellos, consultar el apartado *glmnet()* del manual de Friedman et al. (2023).

Como resultados principales, se pueden obtener los coeficientes del modelo ajustado para cada uno de los *lambda* considerados en la función *glmnet()* aplicando sobre la orden anterior la función *coef()* o un breve resumen con la orden *print()* en el que para cada *lambda* se puede consultar el número de coeficientes no nulos en el modelo (*Df*) y el porcentaje de desviación explicada (*%Dev*). Además, de manera gráfica se puede estudiar la evolución de cada uno de los parámetros según el valor del coeficiente de penalización *lambda* con la orden

$$plot(glmnet(x,\ y,\ family\ =\ "cox",...)).$$

[15]La técnica de máxima verosimilitud penalizada es análoga a la técnica de máxima verosimilitud tradicional pero teniendo en cuenta la presencia de una penalización en la función de máxima verosimilitud que se utiliza a la hora de estimar los parámetros del modelo en estudio.

[16]El algoritmo de descenso por coordenadas cíclicas es una variante del algoritmo de optimización de descenso por coordenadas. El algoritmo en estudio se basa en el hecho de que cada iteración se realiza sobre las coordenadas del vector $\underline{x}$ de manera cíclica mientras se dejan fijas el resto de coordenadas del problema. Se suele utilizar en problemas de optimización convexa en los que es costoso actualizar todas las coordenadas del problema de manera simultánea.

Dado que con la orden *glmnet()* se obtienen diferentes modelos de Cox en función del coeficiente de penalización *lambda* considerado, es de interés seleccionar aquel *lambda* que proporcione un modelo óptimo según diferentes criterios. Por ejemplo, es posible el estudio del *lambda* óptimo con el uso de la validación cruzada. Para ello, se puede usar la función *cv.glmnet()* del propio paquete *glmnet*. La estructura de dicha función es parecida a la de *glmnet* y sus argumentos son:

- x: matriz que contiene para cada una de las observaciones en estudio (filas) el valor asociado de las variables explicativas contempladas en el modelo de Cox a ajustar.
- y: matriz de dos columnas que contiene para cada una de las observaciones en estudio el valor t junto con el estado de censura de cada una de ellas u objeto de tipo *Surv*.
- *family*: cadena de caracteres que representa la familia a la que pertenece el modelo que se quiere ajustar. Puede tomar los mismos valores que se enumeraron en la función *glmnet()*.
- *lambda*: secuencia de valores del coeficiente de penalización asociado al ajuste del modelo a construir.
- *type.measure*: tipo de medida a utilizar como criterio de selección del *lambda* óptimo. Puede tomar diferentes valores pero para el caso del modelo de Cox, los valores posibles son *"deviance"* (basado en la verosimilitud parcial y es el valor por defecto) o *"C"* (índice C de Harrell[17]).
- *nfolds*: número de pliegues a utilizar en el proceso de validación cruzada. Por defecto, toma el valor 10.

Para estudiar el valor óptimo de *lambda* usando la función enunciada, se usará la orden *plot(cv.glmnet())* que representa el logaritmo de *lambda* frente a la medida escogida en el argumento *type.measure*. En dicha gráfica, aparecen siempre dos líneas verticales y discontinuas que señalan dos valores óptimos del parámetro *lambda*: el valor correspondiente al error por validación cruzada mínimo y el valor asociado a la desviación estándar del mínimo (1SE). Para obtener numéricamente dichos valores, se utilizan las órdenes $cv.glmnet()\$lambda.min$ y $cv.glmnet()\$lambda.1se$, respectivamente. Una vez seleccionados los coeficientes, se vuelve a repetir la orden $glmnet()$ con el valor óptimo del *lambda*, de forma que se ajuste el modelo creado.

Por último, destacar que en el caso de que se quieran modificar los parámetros internos de convergencia de la función *glmnet()* como pueden ser el número máximo de iteraciones a la hora de la optimización de los coeficientes de los modelos, se utilizará la orden *glmnet.control()*. Los argumentos de esta función junto con su detalle se pueden consultar en el apartado de dicha orden del manual de Friedman et al. (2023).

[17]El índice de Harrell es una medida de concordancia para los datos de supervivencia comparable con el área bajo la curva (AUC). Toma valores entre 0 y 1, y mide la capacidad predictiva que tiene un modelo de riesgos proporcionales de Cox. Además, dicho índice se considera una medida importante para evaluar la calidad y el ajuste del modelo de Cox en estudio, y será mejor cuanto más cercano al 1 sea.

4.2.4. Otras consideraciones del modelo de Cox

Además de todas las funciones presentadas en los subapartados anteriores, existen otras funciones útiles para realizar un estudio completo del modelo de Cox creado. Entre ellas, se encuentra la orden *predict()* de la librería *survival*, la cual permite obtener las predicciones sobre un nuevo conjunto de datos que se le proporcione al programa. La estructura de la orden es

predict(object, newdata, type, se.fit),

cuyos argumentos se definen como:

- *object*: objeto *coxph* que contiene el modelo de Cox creado.
- *newdata*: nuevo conjunto de datos sobre el que se quiere obtener las estimaciones usando el modelo de Cox creado en el argumento anterior *object*. En el caso de que no se le proporcione ningún conjunto de datos, proporcionará la predicción del conjunto de datos introducido para el ajuste del modelo.
- *type*: indicador del tipo de valor a predecir. Puede tomar diferentes valores como *"lp"* (predictor lineal), *"risk"* (exponencial del predictor lineal o puntuación de riesgo), *"expected"* (número esperado de eventos dadas las covariantes y el tiempo de seguimiento) o *"survival"* (exponencial de la predicción *"expected"* que se corresponde con la probabilidad de supervivencia de un individuo cambiada de signo).
- *se.fit*: valor lógico que muestra el error estándar puntual asociado a cada una de las predicciones realizadas en el caso de tomar el valor *TRUE*. Por defecto, toma el valor *FALSE*.

Como salida, se obtiene la predicción realizada (*fit*) y el error estándar puntual (*se.fit*) asociado a la predicción. Para un estudio más detallado de la orden *predict()*, consultar el apartado de *predict.coxph* del manual de Therneau (2023).

Finalmente, una vez ajustado, evaluado y comprobado que un modelo de Cox es válido, es interesante el estudio de la distribución de la función de supervivencia $S(t)$ asociada al modelo de riesgos proporcionales creado. Para esta finalidad, se usará la orden

survfit(formula, newdata, se.fit, conf.int, stype, ctype, conf.type, censor, start.time, id,...)

de la librería *survival*. Los argumentos de dicha orden se detallan a continuación:

- *formula*: objeto de tipo *coxph* que contiene un modelo de Cox.
- *newdata*: nuevo conjunto de datos (los datos deben tener los mismos nombres de variables que el conjunto de datos original que se usa para la creación del modelo de Cox) para el cual se va a realizar la representación de la curva de supervivencia en función del modelo de Cox introducido.
- *se.fit*: valor lógico que indica si se deben calcular los errores estándar. El valor predeterminado es *TRUE*.
- *conf.int*: indica el nivel de confianza a utilizar para la creación de los intervalos de confianza bilaterales. Por defecto, el valor que tiene es 0.95.

- *stype*: tipo de cálculo de la curva de supervivencia. Puede tomar los valores 1 (cálculo de la curva de supervivencia) o 2 (cálculo de la exponencial del riesgo acumulado). Por defecto, toma el valor 2.
- *ctype*: indica si el cálculo de la tasa de riesgo acumulativo debe calcularse con una corrección. Los valores posibles son 1 (no) o 2 (sí). Por defecto, toma el valor 2.
- *conf.type*: indicador del tipo de estimación del intervalo a realizar. Los posibles valores son *"log"* (valor por defecto y genera intervalos basados en la tasa de riesgo acumulada o el logaritmo de la curva de supervivencia) *"log-log"* (genera el logaritmo de la opción *"log"*), *"plain"* (genera intervalos de tipo estándar), *"none"* (no genera intervalos de confianza) o *"logit"* (genera el logaritmo del cociente de la curva de supervivencia entre la unidad menos la curva de supervivencia).
- *censor*: valor lógico que indica si se incluyen o no en el resultado las observaciones sobre las que no se ha producido el evento en estudio. Por defecto, toma el valor *TRUE*.

La estimación de la función de supervivencia se puede consultar con la orden *summary()* y la información que devuelve es la siguiente:

- *time*: instante de la observación o tiempos observados.
- *n.risk*: número de individuos en riesgo en cada instante de tiempo.
- *n.event*: número de individuos que presentan el evento en cada instante de tiempo.
- *survival*: estimación de la función de supervivencia.
- *std.err*: desviación estándar de la estimación realizada.
- *lower 95 % CI* y *upper 95 % CI*: extremo inferior y superior del intervalo de confianza de la estimación al 95 %.

Además, se puede consultar también el número de observaciones censuradas en cada instante de tiempo en estudio (*n.censor*). Para ampliar la información de manera más detallada, consultar los argumentos de esta orden en el apartado *survfit.coxph* del manual de Therneau (2023).

De manera gráfica, se puede estudiar la estimación de la función de supervivencia calculada junto con el intervalo de confianza obtenido para dicha estimación. La mencionada gráfica se puede realizar con la ayuda del paquete *survminer* (Biecek et al. (2021)) y la función *ggsurvplot()*.

4.3. Aplicaciones

En los próximos dos subapartados se pondrán en práctica las funciones y órdenes descritas en el apartado anterior. En el primero de ellos, las funciones se aplicarán a un conjunto de datos llamado *uissurv.csv* extraído de la asignatura de Modelado y Predicción Estadística y que está formado por variables independientes del tiempo. Dicho conjunto de datos

permitirá mostrar el ajuste de un modelo de Cox con ese tipo de variables, las cuales se presentan y describen en el siguiente subapartado. Por otro lado, en el segundo de los subapartados se mostrará la forma con la que se debe trabajar cuando el conjunto de datos posee covariantes dependientes del tiempo. Para tal caso y con la ayuda de Therneau et al. (2022), se usarán los datos *pbc* y *heart* disponibles en el paquete *survival*. A lo largo del segundo subapartado de esta sección, se mostrarán ambos conjuntos de datos, explicando sus objetivos y detallando las variables de estudio de cada uno de ellos. Además de poder mostrar todo el potencial de las funciones que proporciona R, se mostrarán las conclusiones lógicas que derivan de las salidas obtenidas a lo largo del estudio de los modelos de Cox que aportarán información relevante para la realización de un análisis completo de dichos modelos.

4.3.1. Modelo de Cox con covariantes independientes del tiempo

Tal y como se ha presentado en la introducción anterior, para la construcción y análisis de un modelo de Cox con covariantes independientes del tiempo se usará el conjunto de datos procedente del archivo *uissurv.csv*. Dicho fichero posee una serie de datos procedentes de un estudio titulado *UMARU Impact Study (UIS)*, resultado de un proyecto colaborativo de investigación llevado a cabo entre los años 1989 y 1994 por la Unidad de Investigación sobre el SIDA de la Universidad de Massachusetts (UMARU) y dirigido por los doctores Jane McCusker, Carol Bigelow y Anne Stoddard (Lemeshow et al. (2008)). El objetivo del estudio fue comparar dos tratamientos diferentes diseñados para reducir el abuso de drogas y prevenir conductas de alto riesgo de contagio de VIH. Para ello, se diseñaron dos programas diferentes y el UIS buscaba determinar si la eficacia de los tratamientos en estudio dependía de la duración del programa diseñado. A continuación, se enuncian los dos programas contemplados en el estudio:

a) Programa A: con una duración comprendida entre 3 y 6 meses, los 444 individuos sometidos a este programa debían asistir a un curso de educación sanitaria y prevención de recaídas. Durante dicho periodo de tiempo, los individuos aprenden a reconocer situaciones de alto riesgo que podían provocar una recaída y adquirían las habilidades necesarias para no usar las drogas en dichos momentos.

b) Programa B: 184 individuos eran sometidos a un programa que duraba de 6 a 12 meses y en el que aprendían un estilo de vida muy estructurado en un entorno social de comunidad.

Las variables involucradas en el estudio son:

- *id*: identificador de cada uno de los individuos de la muestra considerada en el estudio.
- *age*: edad de cada individuo expresada en años.
- *beck*: puntuación que el sujeto ha obtenido en el test de depresión de Beck (BDI)[18]. Los valores están comprendidos entre 0 y 54.

[18]El test de depresión de Beck es un cuestionario que consta de 21 preguntas de respuesta múltiple utilizado para medir la severidad de una depresión y fue desarrollado por el psiquiatra, investigador y fundador de la terapia cognitiva, Aaron T. Beck.

- *hercoc*: variable categórica que especifica si el sujeto ha usado heroína/cocaína durante los 3 meses previos a la participación en el estudio. Los posibles valores son: 1 = Heroína & Cocaína (Heroin & Cocain), 2 = Solo Heroína (Heroin only), 3 = Solo Cocaína (Cocain only) o 4 = Ni Heroína ni Cocaína (Neither).
- *ivhx*: indicador de consumo de drogas de uso intravenoso (drogas IV) antes de la admisión al programa. Los posibles valores son: 1 = Nunca (Never), 2 = Previamente (Previous) o 3 = Recientemente (Recent).
- *ndrugtx*: número de tratamientos previos a la incorporación al programa. El rango de valores está comprendido entre 0 y 40.
- *race*: raza del individuo. Los posibles valores son: 0 = Blanca (White) o 1 = No blanca (Non-White).
- *treat*: variable que denota el tratamiento aleatorizado asignado a cada uno de los individuos. Los posibles valores son 0 = Corto (Short) o 1 = Largo (Long).
- *site*: tipo de programa asociado a cada uno de los individuos. Los valores contemplados son 0 = A o 1 = B.
- *los*: tiempo de estancia en el tratamiento expresado en días y medido desde el día de admisión hasta el día de salida del programa.
- *time*: tiempo hasta la recaída en las drogas expresado en días y medido desde el día de admisión del individuo en el programa.
- *censor*: estado de la censura (1 = Recaída en las drogas o 0 = censurado).

Una vez presentadas cada una de las variables, se leerá el fichero que contiene dichos datos:

```
uissurv <- read.csv("uissurv.csv", header = TRUE, sep = ";")
```

Además, se mostrará una muestra aleatoria del objeto *uissurv* que será un *data.frame* que contendrá la información necesaria para construir posteriormente un modelo de Cox:

id	*age*	*beck*	*hercoc*	*ivhx*	*ndrugtx*	*race*	*treat*	*site*	*los*	*time*	*censor*
284	33	17.00	Cocain only	Never	3	Non-White	Short	A	55	115	1
623	36	9.00	Neither	Previous	NA	White	Short	B	37	51	1
98	26	11.00	Heroin & Cocain	Recent	1	White	Long	A	40	73	1
104	25	6.00	Heroin only	Recent	5	White	Long	A	90	655	0
5	24	5.00	Heroin only	Never	5	Non-White	Long	A	173	551	0
33	37	9.45	Neither	Recent	1	White	Short	A	90	259	1
454	43	9.00	Heroin & Cocain	Recent	0	Non-White	Long	B	6	6	1
499	30	8.40	Cocain only	Previous	40	White	Short	B	36	36	1

Tabla 4.1: Muestra de datos del fichero *uissurv*.

Una vez comprendida la estructura de los datos, se procede a construir un modelo de Cox para la variable tiempo de supervivencia frente al resto de variables que serán las variables explicativas del modelo:

```
library(survival)
mod_cox_indp_t <- coxph(
  Surv(time, censor) ~ age + beck + hercoc +
    ivhx + ndrugtx + race + treat + site + los,
  data = uissurv
)
```

A continuación, se adjunta la información que recoge el objeto *coxph* anterior:

```
mod_cox_indp_t
```

```
## Call:
## coxph(formula = Surv(time, censor) ~ age + beck + hercoc + ivhx +
##     ndrugtx + race + treat + site + los, data = uissurv)
##
##                              coef  exp(coef)   se(coef)       z       p
## age                    -0.0209245  0.9792929  0.0082448  -2.538 0.01115
## beck                    0.0053053  1.0053194  0.0049233   1.078 0.28122
## hercocHeroin & Cocain   0.0143785  1.0144823  0.1687127   0.085 0.93208
## hercocHeroin only       0.1660367  1.1806165  0.1646092   1.009 0.31313
## hercocNeither           0.0125511  1.0126302  0.1228395   0.102 0.91862
## ivhxPrevious            0.1882694  1.2071587  0.1387700   1.357 0.17488
## ivhxRecent              0.3986325  1.4897860  0.1482168   2.690 0.00716
## ndrugtx                 0.0262922  1.0266409  0.0085909   3.060 0.00221
## raceWhite               0.3063290  1.3584291  0.1160987   2.639 0.00833
## treatShort             -0.1381856  0.8709370  0.0969330  -1.426 0.15399
## siteB                   0.4437656  1.5585651  0.1112148   3.990 6.6e-05
## los                    -0.0094204  0.9906238  0.0008215 -11.467 < 2e-16
##
## Likelihood ratio test=193.3  on 12 df, p=< 2.2e-16
## n= 575, number of events= 464
##    (53 observations deleted due to missingness)
```

Tal y como se especifica en la presentación de la función *coxph()* del Apartado 4.3.1, en la salida anterior se pueden consultar los coeficientes estimados (columna *coef*) de cada una de las variables explicativas consideradas en el modelo y evaluar su significación con la ayuda de la columna p. Es destacable que, dado que en el conjunto de datos considerado existen variables categóricas con k modalidades, se muestra un coeficiente estimado para cada una de las $k-1$ modalidades de la variable en cuestión. A la vista del resultado obtenido, se sabe que el modelo se ha creado a partir de 575 observaciones de las cuales, en 464 de ellas, se ha producido el evento de interés (recaída en las drogas del individuo) y 53 observaciones no se han usado por poseer valores perdidos.

En el caso de que se quiera tener una salida más detallada del modelo creado, se usará la orden *summary()*:

```
summary(mod_cox_indp_t)
```

```
## Call:
## coxph(formula = Surv(time, censor) ~ age + beck + hercoc + ivhx +
##     ndrugtx + race + treat + site + los, data = uissurv)
##
##   n= 575, number of events= 464
##    (53 observations deleted due to missingness)
##
##                             coef  exp(coef)   se(coef)       z Pr(>|z|)
## age                    -0.0209245  0.9792929  0.0082448  -2.538  0.01115 *
## beck                    0.0053053  1.0053194  0.0049233   1.078  0.28122
## hercocHeroin & Cocain   0.0143785  1.0144823  0.1687127   0.085  0.93208
## hercocHeroin only       0.1660367  1.1806165  0.1646092   1.009  0.31313
## hercocNeither           0.0125511  1.0126302  0.1228395   0.102  0.91862
## ivhxPrevious            0.1882694  1.2071587  0.1387700   1.357  0.17488
## ivhxRecent              0.3986325  1.4897860  0.1482168   2.690  0.00716 **
## ndrugtx                 0.0262922  1.0266409  0.0085909   3.060  0.00221 **
## raceWhite               0.3063290  1.3584291  0.1160987   2.639  0.00833 **
## treatShort             -0.1381856  0.8709370  0.0969330  -1.426  0.15399
## siteB                   0.4437656  1.5585651  0.1112148   3.990  6.6e-05 ***
## los                    -0.0094204  0.9906238  0.0008215 -11.467  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## age                       0.9793     1.0211    0.9636    0.9952
## beck                      1.0053     0.9947    0.9957    1.0151
## hercocHeroin & Cocain     1.0145     0.9857    0.7288    1.4121
## hercocHeroin only         1.1806     0.8470    0.8551    1.6301
## hercocNeither             1.0126     0.9875    0.7960    1.2883
## ivhxPrevious              1.2072     0.8284    0.9197    1.5845
## ivhxRecent                1.4898     0.6712    1.1142    1.9920
## ndrugtx                   1.0266     0.9741    1.0095    1.0441
## raceWhite                 1.3584     0.7361    1.0820    1.7055
## treatShort                0.8709     1.1482    0.7202    1.0532
## siteB                     1.5586     0.6416    1.2533    1.9382
## los                       0.9906     1.0095    0.9890    0.9922
##
## Concordance= 0.741  (se = 0.011 )
## Likelihood ratio test= 193.3  on 12 df,   p=<2e-16
## Wald test            = 170.5  on 12 df,   p=<2e-16
## Score (logrank) test = 176.2  on 12 df,   p=<2e-16
```

A lo largo de las siguientes secciones, se irá analizando el modelo de Cox creado con las covariantes independientes del tiempo contenidas en el conjunto de datos bajo estudio. En primer lugar, se realizará la diagnosis del modelo de forma que se pueda ajustar un modelo que no viole las hipótesis de riesgos tanto global como individualmente. Una vez obtenido dicho modelo, se analizarán los parámetros estimados y los contrastes de hipótesis asociados a estos; y se terminará con la evaluación del mismo.

4.3.1.1. Diagnosis del modelo

Para obtener un modelo de Cox válido y con capacidad predictiva es fundamental que las variables involucradas en él cumplan las hipótesis de riesgos proporcionales tanto de

manera conjunta como individual. El estudio de si la hipótesis de riesgos proporcionales se verifica o no en el modelo se puede realizar de manera gráfica y analítica mediante el test de bondad de ajuste (GOF). Tras esta breve introducción, se presentan ambos enfoques y sus correspondientes conclusiones.

A continuación, se procede a analizar los diferentes test de bondad de ajuste asociados a cada una de las variables predictoras del modelo de Cox creado. Para extraer dichos resultados de manera analítica, se usará la orden *cox.zph()* del paquete *survival* con una transformación de los tiempos de supervivencia de rango (*transform = "rank"*). En la siguiente tabla, se puede consultar el resultado obtenido a través de la orden en cuestión. Dicha tabla contiene para cada una de las variables predictoras consideradas en el modelo, el valor del estadístico χ^2 asociado al contraste realizado junto a sus grados de libertad y su p-valor:

```
cox.zph(mod_cox_indp_t, transform = "rank")
```

Variable	*chisq*	*df*	*p*
age	0.397	1	0.5288
beck	2.979	1	0.0843
hercoc	1.179	3	0.7581
ivhx	0.326	2	0.8496
ndrugtx	0.346	1	0.5564
race	0.503	1	0.4781
treat	8.526	1	0.0035
site	2.414	1	0.1202
los	207.403	1	<2e-16
GLOBAL	213.953	12	<2e-16

Tabla 4.2: Estudio de la hipótesis de riesgos proporcionales tanto de manera individual como global para el modelo de Cox creado.

A la vista de los resultados obtenidos, dado que los p-valores asociados a las variables predictoras *treat* y *los* son menores que 0.05, se concluye que existen evidencias significativas para determinar que dichas variables violan la hipótesis de riesgos proporcionales. Es más, el modelo de Cox creado no es válido dado que globalmente se viola también dicha hipótesis. Para intentar solucionar este problema, se deben eliminar dichas variables, replantear el modelo y analizarlo de nuevo.

Otra forma de obtener la tabla anterior es mediante la orden que se adjunta a continuación:

```
cox.zph(mod_cox_indp_t, transform = "rank")$table
```

El estudio analítico de la hipótesis de riegos proporcionales se puede hacer de manera gráfica a través del estudio de los residuos de Schoenfeld tal y como se especificó en el

Apartado 2.5.2. Antes de proceder a hacer las gráficas, se especificará la manera en la que se pueden extraer los residuos de Schoenfeld asociados a cada uno de los individuos de la muestra y a cada variable predictora considerada. Para ello, se usará la función *residuals()* si se quieren obtener los residuos sin escalar y *cox.zph()$y* en caso contrario:

```
residuals(mod_cox_indp_t, type = c("schoenfeld"))
cox.zph(mod_cox_indp_t, transform = "rank")$y
```

time	*age*	*beck*	*hercoc*	*ivhx*	*ndrugtx*	*race*	*treat*	*site*	*los*
4	0.0823	-0.0916	29.973	1.7194	-0.0126	1.259	-2.8373	0.4223	-0.0341
4	0.2277	0.0220	-0.2571	-10.1006	-0.069	3.1586	-2.5264	-1.2056	-0.0264
6	-0.1071	-0.1726	-2.7891	3.3803	-0.0205	-4.7547	2.0453	-0.9294	-0.0194
6	0.2863	-0.1350	-10.6887	14.3498	-0.3027	-4.4125	-3.4116	6.223	-0.0372
7	0.1487	0.057	33.2189	-6.2191	0.4165	0.9228	1.7559	-0.3232	-0.0264
7	-0.15683	0.0809	49.0868	-13.681	0.0496	1.8397	-1.986	-0.9337	-0.0289
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
502	0.0273	0.008	-16.256	1.9374	0.12689	1.2539	-1.259	1.6857	0.0643
516	0.2697	0.0981	-11.126	-1.7983	-0.0675	2.8468	-1.8483	-1.6535	0.0162
519	-0.1844	-0.0224	2.1318	2.330	-0.0213	-4.2965	-1.7249	3.7895	0.0016
559	-0.2312	-0.1011	0.0223	-0.989	0.0421	2.216	-2.338	-0.574	-0.027
568	0.1275	-0.15836	29.0773	-1.8809	0.4715	0.4334	1.7649	-0.0166	-0.0148
659	-0.1485	-0.0033	-2.5585	3.9765	0.0151	-3.5266	-2.348	0.8089	-0.0059

Tabla 4.3: Residuos escalados de Schoenfeld asociados al modelo de Cox creado.

Una vez mostrada la forma en la que se pueden extraer los residuos de Schoenfeld asociados a cada una de las variables explicativas de un modelo de Cox, se procede a realizar unas gráficas de estos para su posterior análisis. Para ello, se puede usar la función *plot(cox.zph())* o directamente la función *ggcoxdiagnostics()* aplicada al objeto *coxph* que contiene el modelo de Cox. La primera de las opciones tiene la peculiaridad de que las gráficas que hace para las variables de tipo factor no se realizan para cada una de las categorías consideradas en el modelo como variables predictoras, sino que hace una única gráfica que resume todas las categorías. En cambio, en el caso de la función *ggcoxdiagnostics()* se dibujan tantas gráficas como coeficientes se han estimado a la hora de ajustar el modelo de Cox creado. A continuación, se presentan los resultados obtenidos con cada una de las opciones estudiadas y las conclusiones obtenidas de ellas:

```
par(mfrow = c(1, 2))
plot(cox.zph(mod_cox_indp_t, transform = "rank"), col = "red")
par(mfrow = c(1, 1))
```

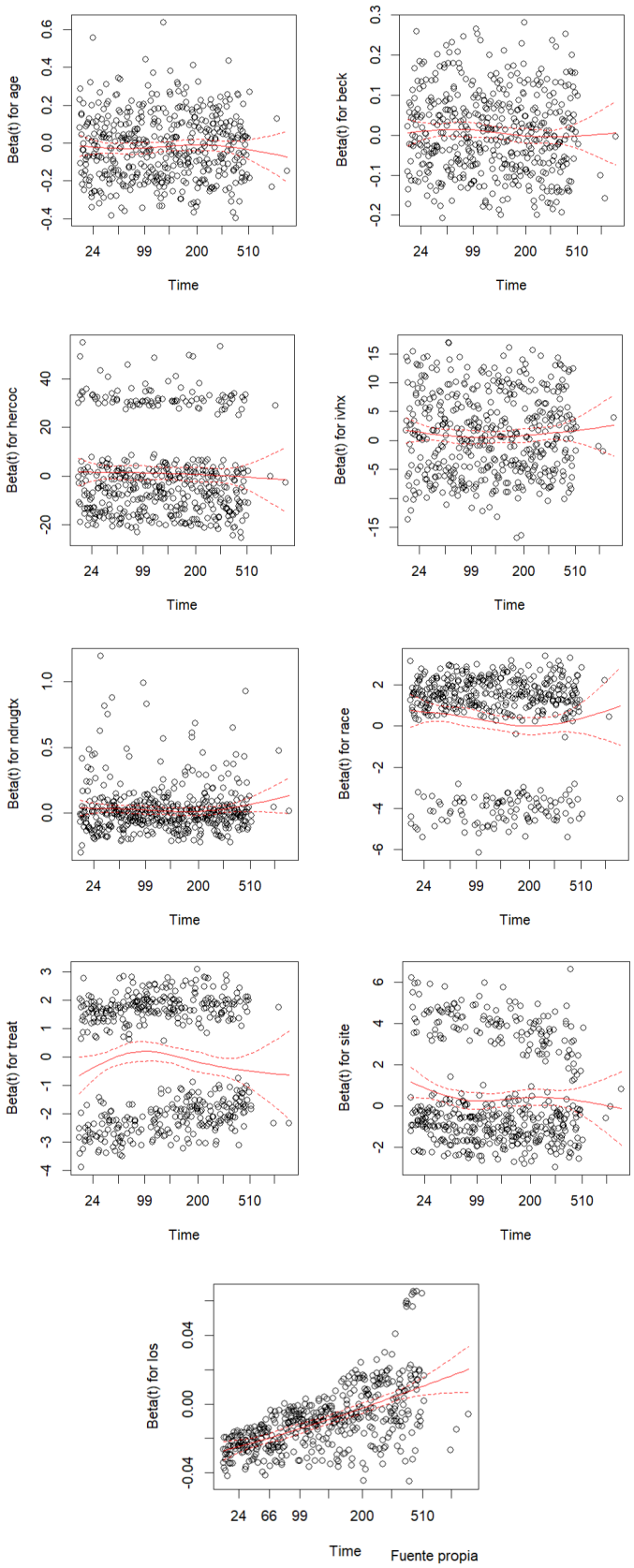
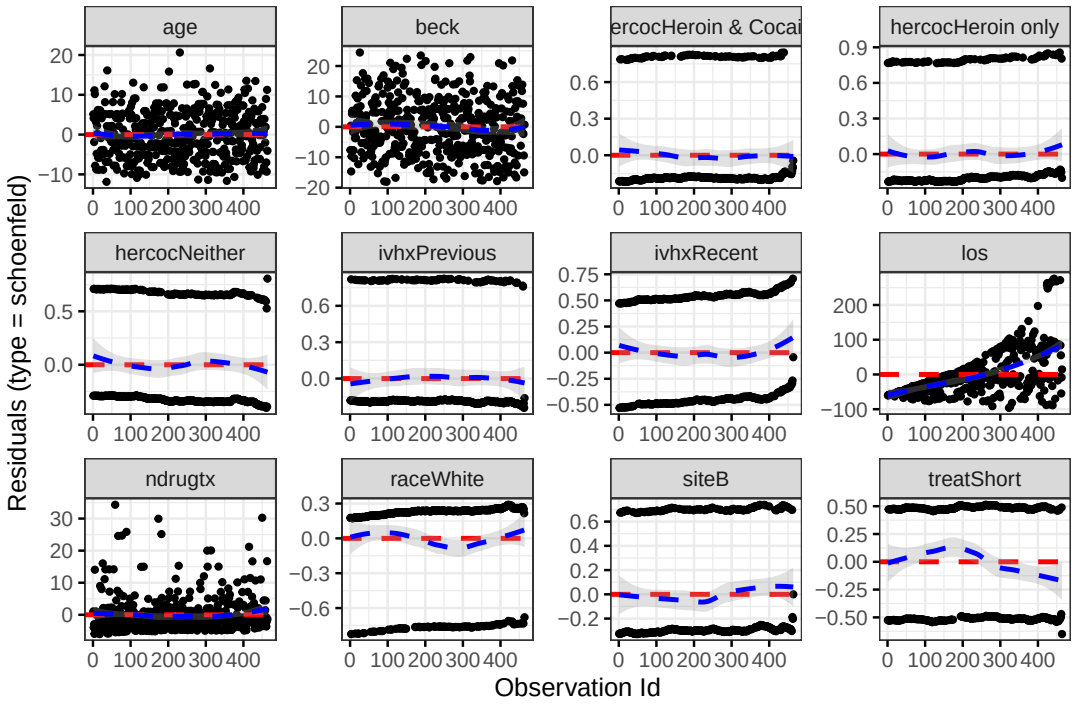


Figura 4.1: Gráficas de los residuos de Schoenfeld con la orden *plot(cox.zph())*.

Tal y como se especificó en el Apartado 2.5.2, se dirá que una variable explicativa del modelo cumple la hipótesis de riesgos proporcionales cuando su gráfica de los residuos frente al tiempo no muestra ningún patrón o, la curva que se dibuja en rojo es de pendiente nula. Teniendo en cuenta dicha apreciación, se podría considerar que efectivamente las variables *treat* y *los* no cumplen la hipótesis de riesgos proporcionales, aunque no de manera muy clara. Dado que las conclusiones de manera gráfica no son muy certeras, es mejor hacer uso de un enfoque más objetivo como son los contrastes de hipótesis.

Tras el análisis gráfico anterior, se procede a mostrar la salida de la segunda de las opciones consideradas para las gráficas de los residuos de Schoenfeld:

```
library(survminer)
ggcoxdiagnostics(mod_cox_indp_t,
  type = "schoenfeld"
)
```



Fuente propia

Figura 4.2: Gráficas de los residuos de Schoenfeld con la orden *ggcoxdiagnostics()* de la librería *survminer*.

Al igual que se ha comentado anteriormente, el análisis de las gráficas de los residuos anteriores para cada una de las variables explicativas consideradas en el modelo no es objetivo por lo que no es aconsejable sacar muchas conclusiones de manera gráfica.

Otra manera de conseguir gráficas de los residuos de Schoenfeld es a través de la función *ggcoxzph()*. Sin embargo, cuando el modelo tiene muchas variables, las gráficas obtenidas no se ven correctamente por lo que solo se adjunta la orden que permite conseguir dichas gráficas:

```
ggcoxzph(cox.zph(mod_cox_indp_t))
```

Asimismo, en el Apartado 2.5.1 se expuso que otra manera gráfica de diagnosticar si una variable cumple o no la hipótesis de riesgos proporcionales es mediante la comparación de las curvas log-log de supervivencia. Este método es útil para las variables categóricas o de tipo factor y se concluirá que dicha variable cumple la hipótesis de riesgos proporcionales cuando las correspondientes curvas son aproximadamente paralelas. A continuación, para mostrar el uso de esta técnica se adjunta un ejemplo realizado sobre la variable *race* y el uso de la función *ggsurvplot()* del paquete *survminer* sobre un objeto de tipo *survfit*:

```
ggsurvplot(survfit(Surv(time, censor) ~ race, data = uissurv),
  fun = "cloglog",
  xlab = "t (días)",
  ylab = "Curva log-log de supervivencia",
  lab = c(10, 10, 7),
  size = 1, legend.title = "Race",
  palette =
    c("chartreuse", "cyan4"),
  conf.int = FALSE,
  legend.labs =
    c("White", "Non-White"), ggtheme = theme_grey(),
  font.caption = c(10, "italic", "black"),
  font.x = c(12, "bold.italic", "black"),
  font.y = c(12, "bold.italic", "black")
)
```

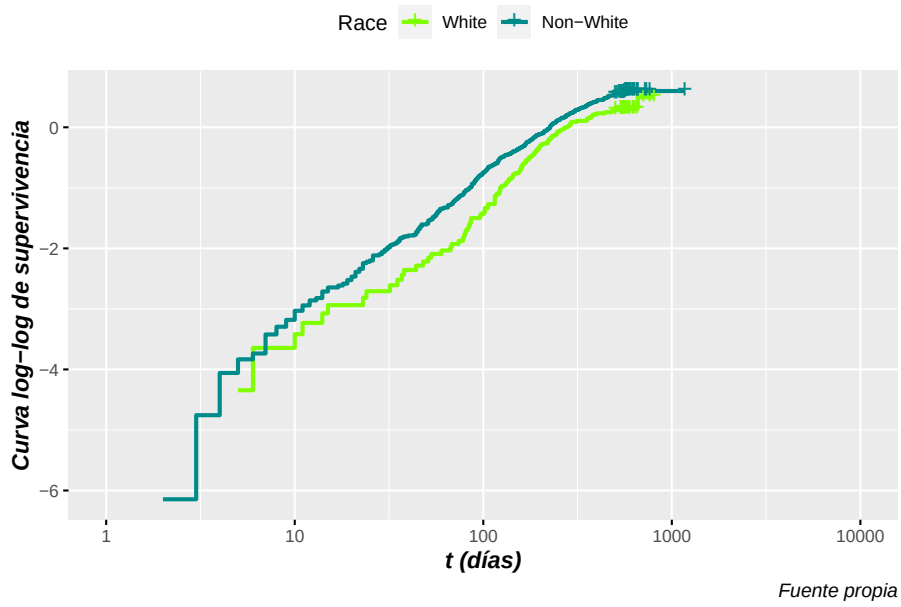


Figura 4.3: Comparación de las curvas log-log de supervivencia para la variable predictora *race*.

A la vista de la gráfica obtenida, se puede razonar que efectivamente, tal y como se concluyó a través del análisis del contraste de bondad de ajuste, la variable *race* cumple la hipótesis de riesgos proporcionales pues además, las curvas log-log de supervivencia para cada uno de los posibles valores de dicha variable (*White* y *Non-White*) son prácticamente paralelas.

Esa misma gráfica pero con una apariencia más sencilla se puede realizar de la siguiente forma con la función *plot()*:

```
plot(survfit(Surv(time, censor) ~ race, data = uissurv),
  fun = "cloglog", xlab = "t(días)",
  ylab = "Curva log-log de supervivencia",
  main = "Curvas log-log de supervivencia por raza"
)
```

Una vez concluido el análisis y comprobado que el modelo creado no es adecuado, se procede a crear un nuevo modelo en el que se eliminan aquellas variables que violan la hipótesis de riesgos proporcionales. De esta forma, el modelo resultante contendrá las variables *age*, *beck*, *hercoc*, *ivhx*, *ndrugtx*, *race* y *site*; y se define a continuación:

```
mod_cox_indp_t_ajus <- coxph(
  Surv(time, censor) ~ age + beck +
    hercoc + ivhx + ndrugtx + race + site,
  data = uissurv
)
```

Una vez definido el modelo, se procede a realizar un estudio que permita concluir si el nuevo modelo creado es válido o no. Para ello, se deberá estudiar de nuevo si las variables involucradas en este nuevo modelo violan o no la hipótesis de riesgos proporcionales de forma individual y global. Este análisis se realizará de manera analítica con el test de bondad de ajuste. En la siguiente tabla, se pueden consultar los p-valores asociados a cada uno de los contrastes realizados a nivel individual y global:

```
cox.zph(mod_cox_indp_t_ajus, transform = "rank")
```

Variable	*chisq*	*df*	*p*
age	0.4774	1	0.49
beck	2.6221	1	0.11
hercoc	0.1591	3	0.98
ivhx	0.2061	2	0.90
ndrugtx	0.5637	1	0.45
race	1.0651	1	0.3

(continúa)

Variable	*chisq*	*df*	*p*
site	0.0512	1	0.82
GLOBAL	5.3431	10	0.87

Tabla 4.4: Estudio de la hipótesis de riesgos proporcionales tanto de manera individual como global para el nuevo modelo de Cox creado.

Dado que todos los p-valores son mayores que 0.05, se puede concluir que no existen evidencias significativas en contra de la hipótesis nula. Por consiguiente, se concluye que todas las variables verifican la hipótesis de riesgos proporcionales a nivel individual y el modelo creado es válido pues el p-valor global también es superior a 0.05.

Una vez encontrado un modelo de Cox en el que todas las variables consideradas verifican la hipótesis de riesgos proporcionales de manera individual y global, se procede a realizar un análisis detallado del modelo con la estimación de parámetros y su significación, así como la evaluación del mismo.

4.3.1.2. Estimación de parámetros

En este apartado, se procederá a analizar los parámetros estimados del modelo y su significación. Sin embargo, antes de realizar dicho análisis, se muestra la información detallada que recoge el objeto *coxph* que contiene el modelo creado:

```
summary(mod_cox_indp_t_ajus)
```

```
## Call:
## coxph(formula = Surv(time, censor) ~ age + beck + hercoc + ivhx +
##     ndrugtx + race + site, data = uissurv)
##
##   n= 575, number of events= 464
##    (53 observations deleted due to missingness)
##
##                             coef exp(coef)  se(coef)      z Pr(>|z|)
## age                    -0.028432  0.971969  0.008136 -3.495 0.000475 ***
## beck                    0.008384  1.008419  0.004952  1.693 0.090414 .
## hercocHeroin & Cocain   0.107710  1.113725  0.164581  0.654 0.512821
## hercocHeroin only       0.147386  1.158802  0.162709  0.906 0.365027
## hercocNeither           0.097513  1.102426  0.121397  0.803 0.421823
## ivhxPrevious            0.178726  1.195693  0.138790  1.288 0.197837
## ivhxRecent              0.292336  1.339553  0.146322  1.998 0.045729 *
## ndrugtx                 0.027330  1.027707  0.008258  3.309 0.000935 ***
## raceWhite               0.215317  1.240255  0.116233  1.852 0.063959 .
## siteB                  -0.076958  0.925929  0.108302 -0.711 0.477341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## age                       0.9720     1.0288    0.9566    0.9876
## beck                      1.0084     0.9917    0.9987    1.0183
## hercocHeroin & Cocain     1.1137     0.8979    0.8067    1.5377
```

```
## hercocHeroin only        1.1588     0.8630    0.8424    1.5941
## hercocNeither            1.1024     0.9071    0.8690    1.3986
## ivhxPrevious             1.1957     0.8363    0.9109    1.5695
## ivhxRecent               1.3396     0.7465    1.0056    1.7845
## ndrugtx                  1.0277     0.9730    1.0112    1.0445
## raceWhite                1.2403     0.8063    0.9876    1.5576
## siteB                    0.9259     1.0800    0.7488    1.1449
##
## Concordance= 0.59  (se = 0.014 )
## Likelihood ratio test= 41.6  on 10 df,   p=9e-06
## Wald test            = 42.66  on 10 df,   p=6e-06
## Score (logrank) test = 43.09  on 10 df,   p=5e-06
```

Una vez obtenida la información del modelo creado, para consultar los parámetros o coeficientes estimados y asociados a cada una de las variables predictoras del modelo, se usará la orden *coxph()$coefficients*:

```
mod_cox_indp_t_ajus$coefficients
```

Variable	*Estimación*
age	-0.028431649
beck	0.008384019
hercocHeroin & Cocain	0.107710247
hercocHeroin only	0.147386411
hercocNeither	0.097513265
ivhxPrevious	0.178725700
ivhxRecent	0.292335842
ndrugtx	0.027329638
raceWhite	0.215317216
siteB	-0.076958222

Tabla 4.5: Estimación de los coeficientes asociados a cada una de las variables explicativas del modelo creado.

El efecto que tiene cada una de las variables sobre el tiempo de supervivencia modelizado se estudia por medio de los efectos multiplicativos sobre la función de riesgo $h(t)$ que se pueden consultar a través de la exponencial de los coeficientes. Esta información se encuentra disponible en el *summary()* del modelo que se presentó anteriormente. A continuación, se adjunta una tabla en la que se pueden consultar dichos coeficientes junto con su intervalo de confianza a un nivel de confianza del 95 %:

```
summary(mod_cox_indp_t_ajus)$conf.int
```

Variable	*exp(coef)*	*lower .95*	*upper .95*
age	0.9719687	0.9565930	0.9875916
beck	1.0084193	0.9986801	1.0182534
hercocHeroin & Cocain	1.1137250	0.8066522	1.5376930
hercocHeroin only	1.1588017	0.8423854	1.5940699
hercocNeither	1.1024261	0.8689935	1.3985642
ivhxPrevious	1.1956927	0.9109214	1.5694889
ivhxRecent	1.3395528	1.0055646	1.7844720
ndrugtx	1.0277065	1.0112063	1.0444760
raceWhite	1.2402553	0.9875832	1.5575732
siteB	0.9259285	0.7488422	1.1448923

Tabla 4.6: Estimación de la exponencial de los coeficientes asociados a cada una de las variables explicativas del modelo creado junto a su intervalo de confianza al 95 %.

Se analizará un par de variables para mostrar el efecto que tienen estas sobre la función de riesgo de recaída en las drogas en el modelo creado:

- Variable *age*: en este caso, se tiene que $\hat{\beta}_1 = -0.02843$ y $e^{\hat{\beta}_1} = 0.97196$, lo que se traduce en que manteniendo constantes el resto de variables involucradas en el modelo de Cox creado, el aumento en un año de la edad de un sujeto supone que la función riesgo de recaída de este en las drogas se reduzca en un factor de 0.97196, que en promedio se corresponde con un 2.8 % aproximadamente, ya que $(1 - 0.97196) \cdot 100\,\% \approx 2.804\,\%$.
- Variable *beck*: se ha obtenido que $\hat{\beta}_2 = 0.00838$ y $e^{\hat{\beta}_2} = 1.00842$ y, como consecuencia de ello, si se mantienen constantes el resto de variables involucradas en el modelo de Cox creado y se aumenta en una unidad la puntuación que el sujeto ha obtenido en el test de depresión de Beck, se tiene que la función riesgo de recaída en las drogas de dicho sujeto, $h_i(t \mid \underline{X}_i)$, aumentará de manera aproximada en un 0.8 %, ya que $(1.00842 - 1) \cdot 100\,\% \approx 0.842\,\%$.

Sin embargo, la interpretación completa y detallada del efecto o influencia que tienen cada una de las variables explicativas del modelo sobre la función de riesgo de recaída en las drogas, debe realizarse por un especialista en la materia o área de investigación de la cuestión tratada y depende del valor y del signo de los coeficientes estimados.

Además, se puede estudiar el intervalo de confianza asociado a cada uno de los parámetros estimados. Para ello, se usará la siguiente orden:

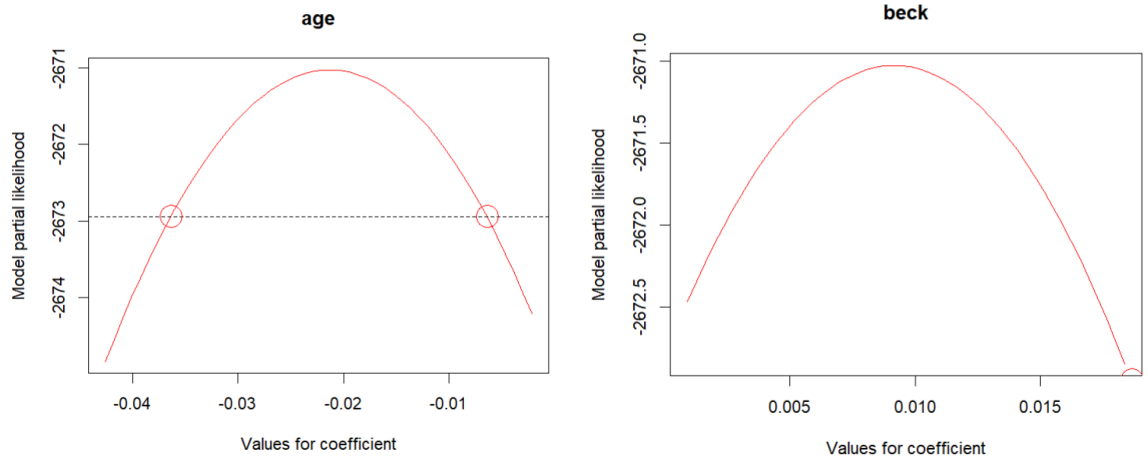
```
confint(mod_cox_indp_t_ajus)
```

Variable	*coef*	*2.5 %*	*97.5 %*
age	-0.028431649	-0.044377267	-0.01248603
beck	0.008384019	-0.001320805	0.01808884
hercocHeroin & Cocain	0.107710247	-0.214862725	0.43028322
hercocHeroin only	0.147386411	-0.171517603	0.46629043
hercocNeither	0.097513265	-0.140419582	0.33544611
ivhxPrevious	0.178725700	-0.093298621	0.45075002
ivhxRecent	0.292335842	0.005549126	0.57912256
ndrugtx	0.027329638	0.011143980	0.04351529
raceWhite	0.215317216	-0.012494566	0.44312900
siteB	-0.076958222	-0.289227050	0.13531061

Tabla 4.7: Estimación de los coeficientes asociados a cada una de las variables explicativas del modelo creado junto a su intervalo de confianza al 95 %.

Una vez presentados los coeficientes estimados del modelo creado, es posible observar el comportamiento gráfico de dichas estimaciones junto con su intervalo de confianza. Para ello, se usará la orden *profLik()* del paquete *survMisc*. Sin embargo, esta orden presenta un pequeño inconveniente, ya que solo funciona para modelos cuyas variables predictoras son todas de tipo numérico. Como consecuencia de ello, no es posible hacer las gráficas asociadas a las variables explicativas del modelo creado. No obstante, para mostrar la salida de esta función, se creará un nuevo modelo que contiene solo las variables numéricas del conjunto de datos en estudio (*age*, *beck* y *ndrugtx*) y se mostrarán las gráficas asociadas a cada uno de los coeficientes estimados:

```
library(survMisc)
profLik(coxph(
  Surv(time, censor) ~ age + beck + ndrugtx,
  data = uissurv
), devNew = F, col = "red")
```



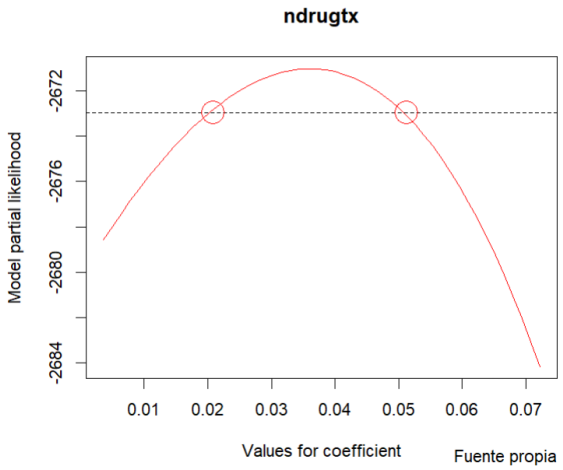


Figura 4.4: Gráficas de los coeficientes estimados obtenidas con la orden *profLik()*.

Por último, es posible consultar el número de iteraciones realizadas en el ajuste para la estimación de los parámetros del modelo con la siguiente orden:

```
mod_cox_indp_t_ajus$iter
```

```
## [1] 4
```

En este estudio, el número de iteraciones que se han necesitado para que el método de estimación de los parámetros converja es 4.

Asimismo, se puede consultar el valor de la log-verosimilitud del modelo (valor de la izquierda) y ese mismo valor bajo la hipótesis nula $H_0 : \underline{\beta} = \underline{0}$ (valor de la derecha):

```
mod_cox_indp_t_ajus$loglik
```

```
## [1] -2663.15 -2642.35
```

4.3.1.3. Contrastes de hipótesis

Una vez analizados los coeficientes del modelo creado, se procede a estudiar si verdaderamente dichos coeficientes son significativos o no para el modelo. Es decir, se debe estudiar y resolver los contrastes de hipótesis individuales siguientes para cada una de las variables explicativas consideradas en el modelo creado; i.e.:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \end{cases}$$

para todo $j = 1, \ldots, 10$. Para ello, se analiza la columna *Pr(>|z|)* del *summary()* que contiene los p-valores asociados a cada una de las estimaciones obtenidas mediante el estadístico de Wald que se adjunta también en la columna z de esa misma orden:

```
summary(mod_cox_indp_t_ajus)$coef[, c(4, 5)]
```

Variable	*z*	*Pr(>|z|)*
age	-3.4946910	0.0004746109 ***
beck	1.6932172	0.0904141405 .
hercocHeroin & Cocain	0.6544510	0.5128212840
hercocHeroin only	0.9058276	0.3650271098
hercocNeither	0.8032623	0.4218231408
ivhxPrevious	1.2877376	0.1978373357
ivhxRecent	1.9978879	0.0457288101 *
ndrugtx	3.3094180	0.0009349015 ***
raceWhite	1.8524678	0.0639586787 .
siteB	-0.7105864	0.4773405668

Tabla 4.8: Contrastes de hipótesis individuales asociados al modelo de Cox creado.

A la vista de los resultados obtenidos, se concluye que las variables significativas del modelo son *age*, *ivhxRecent* y *ndrugtx* dado que el p-valor asociado a ellas es menor que 0.05; y las variables no significativas son *beck*, *hercocHeroin & Cocain*, *hercocHeroin only*, *hercocNeither*, *ivhxPrevious*, *raceWhite* y *siteB*. Estas últimas variables no significativas podrían eliminarse del modelo y este no sufriría ningún cambio. Además, en el caso de que se considerase un nivel de significación del 10 %, también pasarían a ser significativas las variables *beck* y *raceWhite*.

Una vez realizados y analizados los contrastes de hipótesis individuales, es momento de analizar lo que ocurre con el contraste de hipótesis conjunto de todas las covariantes usadas en el modelo:

$$\begin{cases} H_0 : \underline{\beta} = (\beta_{10}, \ldots, \beta_{100})' = \underline{0} \\ H_1 : \beta_j \neq \beta_{j_0} \text{ para algún } j = 1, \ldots, 10. \end{cases}$$

Para dicho cometido, existen tres formas de resolverlo: mediante el test de razón de verosimilitud, el test de Wald o el test Score, los cuales son asintóticamente equivalentes. Tanto el estadístico asociado a cada uno de dichos test como su p-valor, se pueden consultar en el *summary()* del modelo creado y se adjunta a continuación:

```
Likelihood ratio test= 41.6  on 10 df,   p=9e-06
Wald test            = 42.66  on 10 df,   p=6e-06
Score (logrank) test = 43.09  on 10 df,   p=5e-06
```

Figura 4.5: Captura de la salida *summary(mod_cox_indp_t_ajus)* para el estudio del contraste de hipótesis conjunto del modelo de Cox.

Dado que los p-valores obtenidos son todos significativos, i.e., menores que 0.05, se concluye que el modelo creado permite explicar la variable tiempo hasta la recaída en

las drogas. Además, el valor de los estadísticos χ^2 asociados a los test en estudio junto a sus grados de libertad y p-valores se pueden consultar directamente con las siguientes órdenes:

- Test de razón de verosimilitud:

```
summary(mod_cox_indp_t_ajus)$logtest
```

test	*df*	*pvalue*
4.160053e+01	1.000000e+01	8.830632e-06

Tabla 4.9: Test de razón de verosimilitud del modelo de Cox creado.

- Test de Wald:

```
summary(mod_cox_indp_t_ajus)$waldtest
```

test	*df*	*pvalue*
4.26600e+01	1.00000e+01	5.72741e-06

Tabla 4.10: Test de Wald del modelo de Cox creado.

- Test score:

```
summary(mod_cox_indp_t_ajus)$sctest
```

test	*df*	*pvalue*
4.309274e+01	1.000000e+01	4.786700e-06

Tabla 4.11: Test score del modelo de Cox creado.

Por último, es destacable que todo el estudio analítico del modelo que se ha realizado a través de la estimación de los parámetros y de los contrastes de hipótesis individuales y conjunto, se puede resumir gráficamente con la orden *ggforest()* de la librería *survminer*. En dicha salida y, tal y como se especificó en el Apartado 4.3.1, se puede consultar el coeficiente estimado de cada una de las variables explicativas junto con su intervalo de confianza, el número de observaciones asociadas a cada una de ellas o los p-valores asociados a cada uno de los parámetros, entre otra información relevante. La orden en cuestión, se muestra en la siguiente línea de código:

```
ggforest(mod_cox_indp_t_ajus, main = "Resumen del modelo")
```

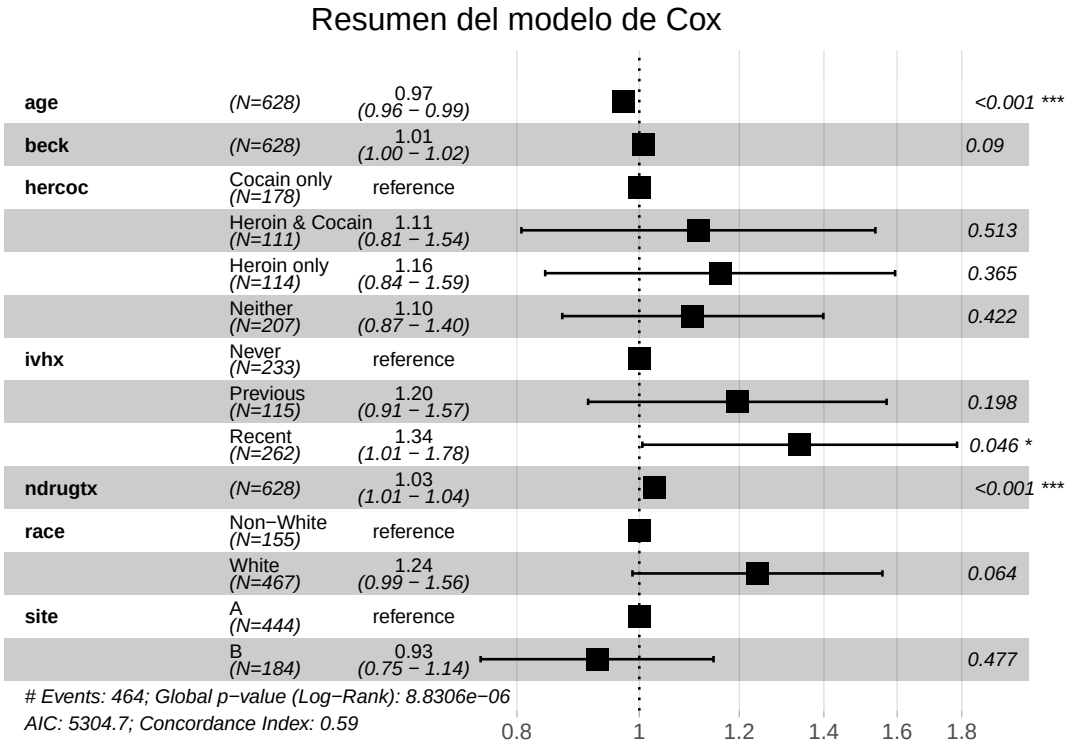


Figura 4.6: Resumen gráfico del análisis del modelo de Cox creado con la función *ggforest()* del paquete *survminer*.

En el resumen anterior, se observan cinco columnas bien diferenciadas: la primera de ellas contiene el nombre de las variables explicativas consideradas en el modelo de Cox creado; la segunda, especifica los diferentes valores o categorías que toman las variables de tipo factor junto con el número de individuos u observaciones que presentan cada una de ellas; en la tercera, se muestra la exponencial del coeficiente estimado de cada variable junto con su intervalo de confianza al 95 % (valor por defecto); en la cuarta, se dibuja dicha estimación con su intervalo de confianza; y por último, en la quinta, se presenta el p-valor asociado a las estimaciones de los parámetros realizadas. Además, en el extremo inferior izquierdo también se puede consultar un resumen global del modelo creado que incluye el p-valor del contraste conjunto o el número de sujetos sobre los que se ha producido el evento de interés. Dado que con esta gráfica se está analizando el mismo modelo de Cox construido al principio de la sección, las conclusiones que se obtienen consultando esta última gráfica son las mismas que se han ido destacando a lo largo de los Apartados 4.3.1.2 y 4.3.1.3.

4.3.1.4. Evaluación del modelo

Para evaluar la calidad de un modelo de Cox, se usarán las medidas expuestas en el Apartado 2.6 que incluyen el coeficiente sugerido por Nagelkerke (1991), el propuesto

por O'Quigley et al. (2005) y el sugerido por Royston (2006). Todos estos coeficientes se pueden obtener con la orden *rsq()* de la librería *survMisc*. A continuación, se adjuntan dichos coeficientes para el nuevo modelo de Cox creado:

```
library(survMisc)
rsq(mod_cox_indp_t_ajus, sigD = NULL)
```

```
## $cod
## [1] 0.06979357
##
## $mer
## [1] 0.08575466
##
## $mev
## [1] 0.05394637
```

En vista a los resultados obtenidos, no parece que la calidad del modelo creado sea excesivamente buena para la predicción del tiempo de supervivencia pues, los coeficientes obtenidos son muy cercanos al 0 e interesaría que fueran próximos a 1. No obstante, el modelo obtenido es válido para determinar aquellas variables que influyen en dicho tiempo y el sentido de dicha influencia.

4.3.1.5. Selección de variables del modelo

En esta sección, se realizará un estudio de selección de variables en un modelo de Cox con la librería *glmnet*. Para llevarlo a cabo, se considerará el modelo de Cox que incluye todas las posibles variables explicativas (excepto la variable *los*) del conjunto de datos *uissurv* con el que se ha estado trabajando en las secciones anteriores. En primer lugar, se hará un estudio de los λ's óptimos ($\lambda_{\text{mín}}$ y λ_{1se}) con la ayuda de la función *cv.glmnet()* y una vez encontrados dichos valores, se crearán los modelos correspondientes con la orden *glmnet()*. Antes de proceder a la programación del modelo, es preciso resaltar que las funciones anteriores operan con estructuras de datos de tipo *matrix* en las que todas sus variables son numéricas y no existen valores perdidos. Debido a esta peculiaridad, es necesario convertir las variables de tipo factor cuyas categorías son cadenas de caracteres en otras variables que toman solo valores numéricas. Como consecuencia de ello y dado que el conjunto de datos *uissurv* a utilizar posee variables de tipo factor sin codificación numérica y con datos perdidos, se procederá a modificarlo para que sea posible su posterior aplicación con las funciones citadas anteriormente.

En primer lugar, se procede a eliminar aquellas observaciones o individuos que poseen valores perdidos en cualquiera de sus variables. El nuevo conjunto de datos se renombrará como *uissurv_glmnet*:

```
uissurv_glmnet <- na.omit(uissurv)
```

Tras eliminar los valores perdidos, se procede a recodificar cada una de las categorías de las variables de tipo factor (*hercoc*, *ivhx*, *race*, *treat* y *site*) para que cumplan el requisito

de ser variables numéricas. Para ello, es necesario el uso de la función *recode()* de la librería *dplyr* (Wickham et al. (2023b)):

```
library(dplyr)
uissurv_glmnet$hercoc <- recode(uissurv_glmnet$hercoc,
  "Heroin & Cocain" = 1,
  "Heroin only" = 2, "Cocain only" = 3,
  "Neither" = 4
)
uissurv_glmnet$ivhx <- recode(uissurv_glmnet$ivhx,
  "Never" = 1,
  "Previous" = 2,
  "Recent" = 3
)
uissurv_glmnet$race <- recode(uissurv_glmnet$race,
  "White" = 0,
  "Non-White" = 1
)

uissurv_glmnet$treat <- recode(uissurv_glmnet$treat,
  "Short" = 0,
  "Long" = 1
)
uissurv_glmnet$site <- recode(uissurv_glmnet$site,
  "A" = 0,
  "B" = 1
)
```

Una vez preparadas las variables categóricas, se procede a crear dos nuevos conjuntos de datos: el primero de ellos, x, contiene las variables explicativas del modelo y el segundo, y, incluye las variables tiempo y censura. Nótese también que para trabajar un modelo de Cox con la librería *glmnet*, es necesario que la columna que contiene los datos de censura se llame *status*. Por lo tanto, será necesario aplicar dicho cambio también al conjunto de datos utilizado. A continuación, se presenta el código necesario para terminar la preparación de los datos:

```
x <- uissurv_glmnet[, c(
  "age", "beck", "hercoc", "ivhx", "ndrugtx",
  "race", "treat", "site"
)]
y <- uissurv_glmnet[, c("time", "censor")]
colnames(y)[2] <- "status"
```

En el momento en el que se tienen los datos preparados, se procede a la construcción del ajuste con la librería *glmnet* y la búsqueda de los λ's óptimos. Recuérdese que para ello, se usará la función *cv.glmnet()* con el argumento *family = "cox"*, *type.measure = "deviance"* para que el ajuste del modelo de Cox sea a través de la verosimilitud parcial, mientras

que el resto de argumentos que toma por defecto ($alpha = 1$[19] y $nlambda = 100$). De esta forma, el estudio que se realizará será una selección de variables por el método Lasso para así conseguir que algunos coeficientes estimados sean nulos. Se empleará también una semilla para que los resultados obtenidos mediante validación cruzada sean reproducibles posteriormente. Asimismo, se añadirá una gráfica en la que se puede observar la evolución del logaritmo del λ utilizado en función de la verosimilitud parcial. A continuación, se procede a realizar el estudio:

```
library(glmnet)
set.seed(1)
cvfit <- cv.glmnet(as.matrix(x), as.matrix(y),
  family = "cox",
  type.measure = "deviance"
)
```

Se adjunta la gráfica en cuestión en la que también pueden consultarse en el extremo superior el número de variables consideradas en los ajustes realizados con los diferentes λ's y con líneas verticales discontinuas se indican los valores óptimos de λ:

```
plot(cvfit)
```

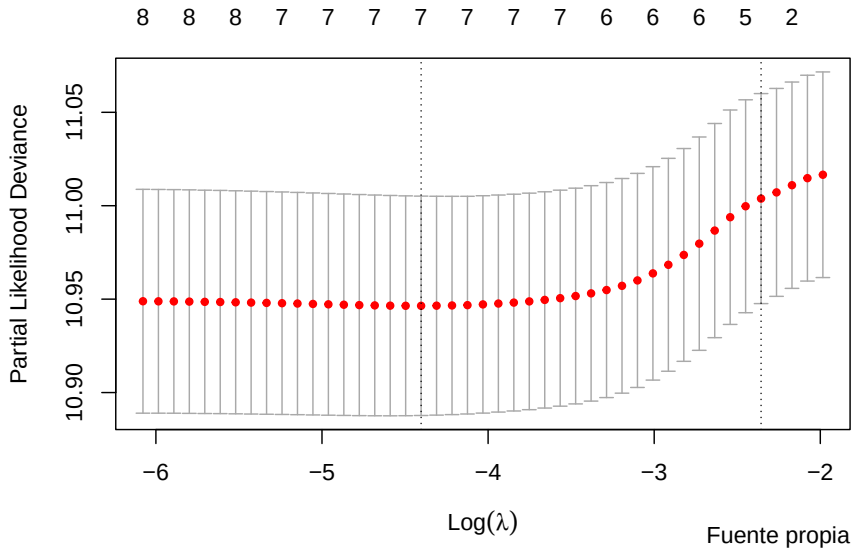


Figura 4.7: Evolución del valor de $log(\lambda)$ en función de la verosimilitud parcial obtenida con la función *cv.glmnet()*.

Se observa que para ambos parámetros óptimos se ha producido una selección de variables, consecuencia directa con el hecho de que se está realizando un ajuste usando el método Lasso. Una vez consultada la gráfica, se extraen los valores óptimos de λ:

[19]El valor del argumento *alpha* determina el método de regularización deseado y toma valores entre 0 y 1. En el caso del valor 0, aplica el método Ridge; y el 1, realiza el método Lasso.

- Valor obtenido de $\lambda_{\text{mín}}$:

```
cvfit$lambda.min
```

```
## [1] 0.01223744
```

- Valor obtenido de λ_{1se}:

```
cvfit$lambda.1se
```

```
## [1] 0.09475009
```

Tras la extracción de los parámetros λ óptimos, se procede a construir los modelos correspondientes con la orden *glmnet()*. El primer modelo se creará con el valor del $\lambda_{\text{mín}}$; y el segundo, con el valor del λ_{1se}:

1. Modelo usando $\lambda_{\text{mín}}$:

```
fit <- glmnet(as.matrix(x), as.matrix(y),
  family = "cox", lambda = cvfit$lambda.min
)
```

Una vez definido el modelo correspondiente con su $\lambda_{\text{mín}}$, se estudia el valor de sus coeficientes estimados:

```
fit$beta
```

Variable	*s0*
age	-0.02421305
beck	0.00664753
hercoc	.
ivhx	0.15320376
ndrugtx	0.02529730
race	-0.17576828
treat	-0.19918352
site	-0.05699427

Tabla 4.12: Estimación de los coeficientes asociados a cada una de las variables explicativas usando $\lambda_{\text{mín}}$ en el método Lasso.

Tal y como se puede observar, con este valor de λ se obtiene un modelo de Cox en el que el coeficiente de la variable *hercoc* es nula. De esta forma, se ha conseguido un nuevo modelo de Cox más simple. Una vez obtenido este modelo, se debería realizar un estudio de diagnosis del modelo para de manera objetiva, saber si el modelo creado es válido o no.

2. Modelo usando λ_{1se}:

```
fit1 <- glmnet(as.matrix(x), as.matrix(y),
  family = "cox", lambda = cvfit$lambda.1se
)
```

Al igual que se ha realizado en el modelo anterior, se estudian cada uno de los coeficientes estimados:

```
fit1$beta
```

Variable	*s0*
age	.
beck	.
hercoc	.
ivhx	0.048330501
ndrugtx	0.006150463
race	0.006891106
treat	.
site	.

Tabla 4.13: Estimación de los coeficientes asociados a cada una de las variables explicativas usando λ_{1se} en el método Lasso.

A la vista de los coeficientes, se ha obtenido un nuevo modelo en el que solo se consideran tres variables explicativas no nulas (*ivhx*, *ndrugtx* y *race*) para modelizar el tiempo hasta la recaída en las drogas. Al igual que se especificó en el caso anterior, para garantizar que el modelo obtenido es válido, será necesario hacer un estudio de diagnosis de dicho modelo. En este caso, el modelo obtenido mediante el método Lasso con λ_{1se} es mucho más simple que el creado con $\lambda_{\text{mín}}$.

Por último, se realiza una gráfica en la que se puede observar la evolución de los valores de los coeficientes de las variables explicativas consideradas frente a la norma L1 a medida que el valor de λ varía. Además, en el eje superior de esta se puede consultar el número de coeficientes distintos de 0:

```
plot(glmnet(as.matrix(x), as.matrix(y), family = "cox"))
```

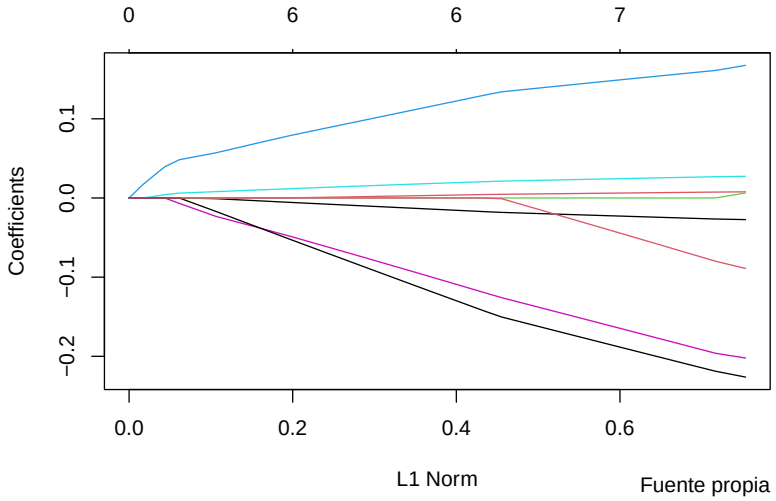


Figura 4.8: Evolución del valor de los coeficientes de las variables predictoras frente a la norma L1 en el método Lasso.

4.3.1.6. Otras consideraciones del modelo

Para terminar con el análisis completo de un modelo de Cox, se añaden órdenes que pueden ser de utilidad a la hora de programar un modelo de regresión de este tipo. A lo largo de esta sección, el modelo que se usará es el *mod_cox_indp_t_ajus* que se ajustó en apartados anteriores. En primer lugar, se muestra la estimación de la matriz de varianzas y covarianzas de las variables predictoras usando la orden *coxph()$var*, la cual permite evaluar la relación existente entre las variables explicativas del modelo:

```
mod_cox_indp_t_ajus$var
```

```
##                [,1]          [,2]          [,3]          [,4]          [,5]
##  [1,]  6.618911e-05  2.423426e-06 -6.345488e-05 -8.855666e-05 -3.738056e-05
##  [2,]  2.423426e-06  2.451767e-05  2.693840e-05  2.899191e-05  4.293907e-05
##  [3,] -6.345488e-05  2.693840e-05  2.708693e-02  1.556921e-02  8.178327e-03
##  [4,] -8.855666e-05  2.899191e-05  1.556921e-02  2.647426e-02  8.290126e-03
##  [5,] -3.738056e-05  4.293907e-05  8.178327e-03  8.290126e-03  1.473712e-02
##  [6,] -2.031930e-04 -5.588681e-05 -1.889368e-03 -1.904810e-03 -2.545384e-03
##  [7,] -2.520241e-04 -7.524165e-05 -1.148694e-02 -1.151066e-02 -1.443276e-03
##  [8,] -1.173992e-05  1.579762e-06  1.749931e-05  6.985647e-05  4.541880e-05
##  [9,]  7.498269e-05  2.613260e-05  9.140565e-04 -8.209515e-04 -1.366308e-04
## [10,]  1.381050e-05  2.908462e-05  7.206944e-04  5.188565e-04 -4.136942e-04
##                [,6]          [,7]          [,8]          [,9]         [,10]
##  [1,] -2.031930e-04 -2.520241e-04 -1.173992e-05  7.498269e-05  1.381050e-05
##  [2,] -5.588681e-05 -7.524165e-05  1.579762e-06  2.613260e-05  2.908462e-05
##  [3,] -1.889368e-03 -1.148694e-02  1.749931e-05  9.140565e-04  7.206944e-04
##  [4,] -1.904810e-03 -1.151066e-02  6.985647e-05 -8.209515e-04  5.188565e-04
##  [5,] -2.545384e-03 -1.443276e-03  4.541880e-05 -1.366308e-04 -4.136942e-04
##  [6,]  1.926279e-02  8.321383e-03 -1.205006e-04 -1.520755e-03 -1.095199e-03
##  [7,]  8.321383e-03  2.141026e-02 -2.283340e-04 -3.168170e-03  1.992492e-03
##  [8,] -1.205006e-04 -2.283340e-04  6.819688e-05 -6.367003e-05  1.111069e-05
```

```
##  [9,] -1.520755e-03 -3.168170e-03 -6.367003e-05  1.351003e-02 -2.506650e-03
## [10,] -1.095199e-03  1.992492e-03  1.111069e-05 -2.506650e-03  1.172941e-02
```

Cuando un modelo está ajustado es posible predecir el tiempo hasta la recaída en las drogas de nuevas observaciones y para ello, se usará la orden *predict()*. Nótese que para predecir, ajustar y evaluar el rendimiento del modelo, así como evitar el sobreajuste, hubiera sido interesante haber considerado dos conjuntos de datos, el de entrenamiento y el test. A pesar de ello, se procede a hacer una predicción sobre un conjunto de datos generado aleatoriamente:

```
set.seed(12)
age <- sample(18:60, 10, replace = TRUE)
beck <- sample(0:54, 10, replace = TRUE)
hercoc <- sample(c(
  "Heroin & Cocain", "Heroin only",
  "Cocain only", "Neither"
), 10, replace = T)
ivhx <- sample(c("Never", "Recent", "Previous"), 10, replace = T)
ndrugtx <- sample(0:40, 10, replace = TRUE)
race <- sample(c("Non-White", "White"), 10, replace = T)
site <- sample(c("A", "B"), 10, replace = T)

datos_pred <- data.frame(
  age, beck, hercoc, ivhx, ndrugtx, race, site
)
```

Una vez generados los datos, se procede a la predicción de estos:

```
predict(mod_cox_indp_t_ajus, newdata = datos_pred, type = "lp")
```

Observación	*Predicción*
1	0.6017720
2	0.3122243
3	1.3195960
4	0.5770842
5	0.9976317
6	1.2379449
7	0.5594878
8	0.6526373
9	0.8530893
10	0.7766950

Tabla 4.14: Predicción sobre un nuevo conjunto de datos generado de manera aleatoria.

En estudios de análisis de supervivencia es usual el estudio de la distribución de la función de supervivencia $S(t)$, y en este caso, se hará lo propio para el modelo de riesgos proporcionales creado usando la estimación de Kaplan-Meier. Para ello, se hará uso de la orden *survfit()* de la librería *survival* tal y como se presenta en el siguiente código:

```
func_superv <- survfit(mod_cox_indp_t_ajus, type = "kaplan-meier")
```

Para conocer la distribución de $S(t)$ se hará un *summary()* del objeto creado en el que se puede consultar para cada tiempo observado: el número de individuos en riesgo (*n.risk*), el número de individuos que presentan el evento (*n.event*), la estimación de $S(t)$ (*survival*) junto con su intervalo de confianza al 95 % (*lower 95 % CI* y *upper 95 % CI*) y la desviación estándar de la estimación realizada (*std.err*):

```
summary(func_superv)
```

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower 95 % CI*	*upper 95 % CI*
4	575	2	0.998	0.00163	0.995	1.000
6	573	2	0.995	0.00234	0.991	1.000
7	571	4	0.991	0.00339	0.984	0.998
8	567	1	0.990	0.00362	0.983	0.997
9	566	1	0.989	0.00384	0.981	0.996
10	565	3	0.985	0.00446	0.976	0.994
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
502	116	1	0.324	0.04670	0.244	0.430
516	111	1	0.322	0.04665	0.242	0.427
519	110	1	0.319	0.04660	0.240	0.425
559	50	1	0.314	0.04661	0.235	0.420
568	37	1	0.308	0.04671	0.229	0.415
659	8	1	0.273	0.05466	0.184	0.404

Tabla 4.15: Estimación de la función de supervivencia a partir del modelo de Cox creado.

Otra forma de obtener la estimación de la función de supervivencia $S(t)$ es mediante la función *surv_adjustedcurves()*. Sin embargo, la información que proporciona esta es mucho menor que la que se obtiene con la orden anterior, ya que solo presenta para cada uno de los tiempos observados su estimación de la función de supervivencia. En contraposición, esta función es útil cuando se pretende estudiar la distribución $S(t)$ para una variable predictora del modelo creado según sus posibles valores. Por ejemplo, se adjunta la estimación de la variable *site* para cada uno de sus posibles valores:

```
surv_adjustedcurves(mod_cox_indp_t_ajus, variable = "site")
```

time	*variable*	*surv*
0	A	1.0000000
4	A	0.9964503
6	A	0.9929037
7	A	0.9858189
8	A	0.9840476
9	A	0.9822761
10	A	0.9769644
.	.	.
.	.	.
.	.	.
502	A	0.1970014
516	A	0.1951849
519	A	0.1933703
559	A	0.1895206
568	A	0.1844969
659	A	0.1583945

Tabla 4.16: Estimación de la función de supervivencia para el programa 'A' de la variable *site*.

time	*variable*	*surv*
0	B	1.0000000
4	B	0.9968293
6	B	0.9934274
7	B	0.9868616
8	B	0.9852194
9	B	0.9835769
10	B	0.9786504
.	.	.
.	.	.
.	.	.
502	B	0.2205210
516	B	0.2186204
519	B	0.2167206
559	B	0.2126861

(continúa)

time	*variable*	*surv*
568	B	0.2074125
659	B	0.1798440

Tabla 4.17: Estimación de la función de supervivencia para el programa 'B' de la variable *site*.

Por último, para completar el estudio de la función de supervivencia, se hace una gráfica de su estimación para todo el conjunto de datos involucrado en el modelo de Cox creado:

```
ggsurvplot(survfit(mod_cox_indp_t),
  data = uissurv, conf.int = T,
  xlab = "t (días)", ylab = "S(t) estimada",
  legend.title = "Estimación",
  legend.labs = "Kaplan-Meier",
  ggtheme = theme_grey()
)
```

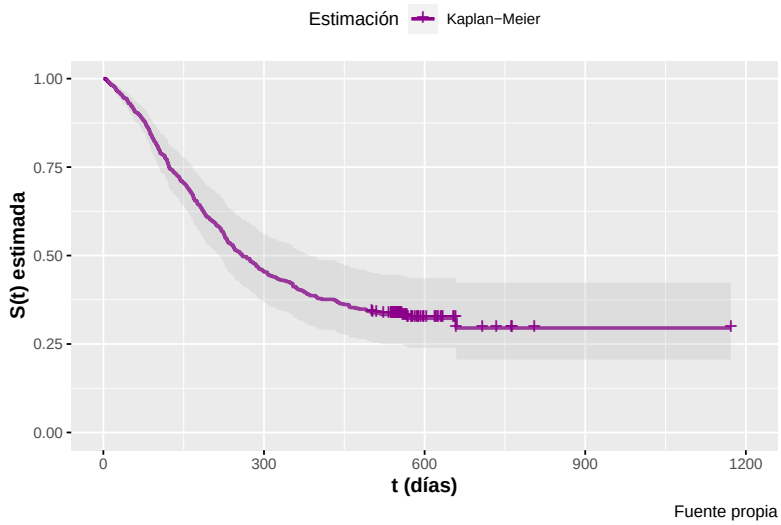


Figura 4.9: Estimación de la función de supervivencia para el modelo creado y ajustado.

La gráfica anterior también puede ejecutarse con la función *ggadjustedcurves()*:

```
ggadjustedcurves(mod_cox_indp_t_ajus,
  xlab = "t (días)",
  ylab = "S(t) estimada",
  legend.title = "Estimación"
)
```

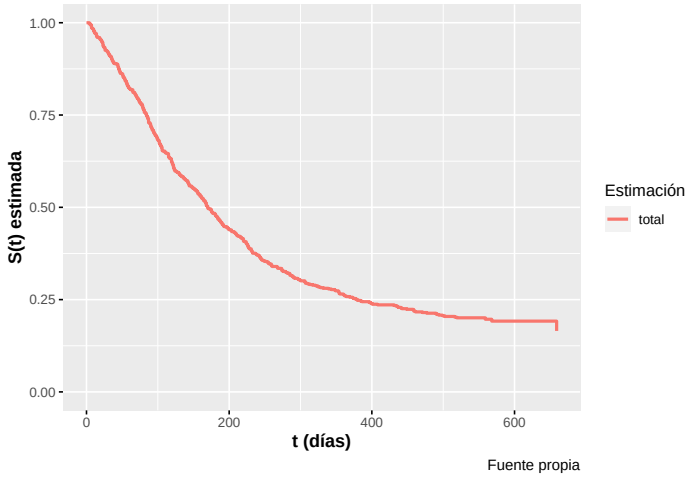


Figura 4.10: Estimación de la función de supervivencia para el modelo creado y ajustado.

Además, la función anterior permite realizar gráficas de las estimaciones de las curvas de supervivencia de cada una de las variables de manera individual. Como ejemplo, se presentan las curvas de supervivencia de las categorías que toma la variable *site* del modelo creado:

```
ggadjustedcurves(mod_cox_indp_t_ajus,
  variable = "site",
  xlab = "t (días)",
  ylab = "S(t) estimada",
  legend.title = "Estimación"
)
```

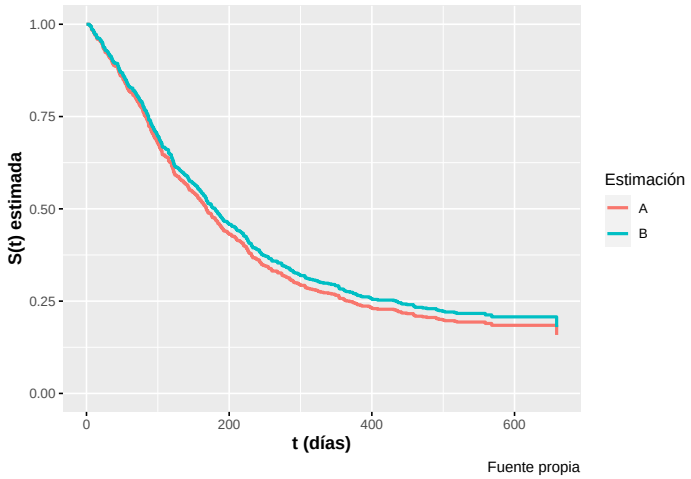


Figura 4.11: Estimación de la función de supervivencia para cada uno de los valores de la variable *site* del modelo creado y ajustado.

4.3.2. Modelo de Cox con covariantes dependientes del tiempo

Al igual que en el apartado anterior se ha realizado un análisis de un modelo de Cox con covariantes independientes del tiempo, en este apartado se hará lo mismo pero tratando además, covariantes dependientes del tiempo. Esta casuística conlleva una complicación adicional en el hecho de que para un único individuo de la muestra o conjunto de datos con el que se trabajará, existirán diferentes filas u observaciones en las que el valor de las variables pueden ir cambiando según el intervalo de tiempo en estudio. Para solucionar esta dificultad cuando los datos no están bien preparados, se hará uso de la función *tmerge()* de la librería *survival* que es capaz de generar un nuevo conjunto de datos que recoge los cambios de valores de las variables para un mismo individuo a lo largo del tiempo. Esta función posee la siguiente estructura:

$$tmerge(data1,\ data2,\ id,\ \ldots,\ tstart,\ tstop,\ \ldots),$$

donde:

- *data1*: conjunto de datos base que se va a completar.
- *data2*: conjunto de datos adicional.
- *id*: identificador en el nuevo conjunto de datos.
- *...*: operaciones que añaden nuevas variables o intervalos al conjunto de datos.
- *tstart*: inicio del intervalo de tiempo en el que se realiza el estudio para cada uno de los individuos.
- *tstop*: fin del intervalo de tiempo en el que se realiza el estudio para cada uno de los individuos.

De esta forma, el nuevo conjunto de datos creado tiene al menos tres nuevas variables que son *id*, *tstart* y *tstop*. No obstante, la clave de la función son los argumentos '*...*' que pueden ser de cuatro tipos:

a) *event()*: este argumento divide el intervalo de estudio original en los diferentes subintervalos donde se producen los eventos o cambios de valores de una o varias variables.

b) *tdc()*: este argumento permite crear una nueva covariable dependiente del tiempo asignándole el valor que toma dicha variable en cada subintervalo de tiempo.

c) *cumevent()*: versión acumulativa de *event()*.

d) *cumtdc()*: versión acumulativa *tdc()*.

Los tiempos asociados a las covariantes dependientes del tiempo se expresan mediante un intervalo abierto a la izquierda y cerrado a la derecha, momento en el cual se produce el evento o una variable cambia su valor.

Una vez que se ha presentado cómo conseguir estructurar correctamente los conjuntos de datos con covariantes dependientes del tiempo que no están bien diseñados, se procede a definir el conjunto de datos *pbc* (Therneau y Grambsch (2000)) de la librería *survival*.

Dicho conjunto de datos contiene la información de un estudio sobre la cirrosis primaria biliar. A su vez, el conjunto *pbcseq* contiene mediciones varias sobre los mismos individuos. Estas mediciones se realizan a lo largo de 10 años de estudio, de modo que, un mismo individuo puede tener diferentes observaciones en el periodo de tiempo estudiado. El objetivo final es la creación de una única base de datos que contenga la información de los conjuntos de datos *pbc* y *pbcseq*. Para ello, es necesario el uso de la función *tmerge()*.

En ambos conjuntos de datos existen variables que identifican a cada uno de los 424 pacientes de la muestra como son su edad (*age*) o sexo (*sex*). Nótese que de esos 424 pacientes, solo se posee información completa de los primeros 312. Además de esas variables identificativas, existen otras muchas variables de naturaleza médica usadas para estudiar la evolución de la enfermedad en cada uno de los pacientes tales como el número de plaquetas (*platalet*) o la cantidad de bilirrubina en el organismo (*bili*). Por último, también contienen información útil para un estudio de análisis de supervivencia: número de días transcurridos desde que el paciente entra en el estudio y su fallecimiento o salida del mismo (*time*); o estado de censura que puede tomar los valores 0, 1 o 2 según el paciente sea censurado, haya sido transplantado o haya muerto. Asimismo, en el conjunto de datos *pbcseq* se encuentra la variable *day* que indica los días en los que se ha realizado las mediciones de las variables médicas anteriormente expuestas.

Una vez presentada la estructura básica de los datos, se procede a unificarlos en un único conjunto de datos que presenta a su vez tanto covariantes dependientes del tiempo (*ascites* (presencia de ascitis), *bili* (cantidad de bilirrubina), *albumin* (cantidad de albúmina), *protime* (tiempo de coagulación sanguínea) y *alk.phos* (cantidad de fosfatasa alcalina)) como independientes del mismo:

1) Se extraen las 312 observaciones completas del conjunto de datos *pbc*:

```
pbc_comp <- subset(pbc, id <= 312)
```

2) Utilizando la función *tmerge()*, se crea una nueva variable *death* con el argumento *event* que marca el cambio de la variable *status*:

```
pbc_unif <- tmerge(pbc_comp,
  pbc_comp,
  id = id,
  death = event(time, status)
)
```

3) Usando de nuevo la función *tmerge()* y el argumento *tdc*, se añade la información del conjunto de datos *pbcseq* de las variables que cambian a lo largo del tiempo:

```
pbc_unif <- tmerge(pbc_unif, pbcseq,
  id = id, ascites = tdc(day, ascites),
  bili = tdc(day, bili), albumin = tdc(day, albumin),
  protime = tdc(day, protime), alk.phos = tdc(day, alk.phos)
)
```

Al aplicar los argumentos *tdc* para crear las nuevas covariantes dependientes del tiempo, se han generado además dos variables nuevas, *tstart* y *tstop*, a partir de las variables *time* y *day*. En dichas variables se encuentra la información de los nuevos intervalos de estudio donde varían las covariantes dependientes del tiempo y cuya estructura es de la forma $(tstart, tstop]$.

De esta forma, se ha obtenido un único conjunto de datos con la información necesaria para realizar un estudio de análisis de supervivencia con algunas variables dependientes del tiempo. A continuación, se adjunta una muestra del conjunto de datos creado donde la variable *id* indica el paciente al que corresponden las mediciones realizadas y donde se han seleccionado las variables más relevantes a lo largo de la construcción del nuevo conjunto de datos:

id	*status*	*ascites*	*bili*	*albumin*	*protime*	*alk.phos*	*tstart*	*tstop*	*death*
1	2	1	14.5	2.60	12.2	1718	0	192	0
1	2	1	21.3	2.94	11.2	1612	192	400	2
84	0	0	0.4	3.76	11.2	1345	0	184	0
84	0	0	0.6	3.6	10.5	784	184	371	0
84	0	0	0.4	3.62	10.4	771	371	4032	0

Tabla 4.18: Muestra de datos del conjunto creado *pbc_unif* con la función *tmerge()*.

Dado que el conjunto de datos creado con la función *tmerge()* es complejo y posee muchas variables, se ha estimado más conveniente usar el conjunto de datos *heart* (Crowley y Hu (1977)) de la librería *survival* para la construcción de un modelo de Cox con covariantes dependientes del tiempo. Dicho conjunto de datos estudia la supervivencia de los pacientes en lista de espera del programa de trasplantes cardíacos de la ciudad de Stanford y posee las siguientes variables:

- *start* y *stop*: tiempo de inicio y fin del intervalo de tiempo en estudio.
- *event*: estado en el intervalo asociado. Puede tomar los valores 0 = vivo y 1 = muerto.
- *age*: edad del paciente menos 48 años.
- *year*: año de aceptación en el programa. Medida expresada en años.
- *surgery*: variable que indica si el paciente ha tenido una cirugía de bypass previa o no. Los posibles valores son: 1 si el paciente ha tenido cirugía previa o 0 en caso contrario. Se trata de una variable dependiente del tiempo.
- *transplant*: variable binaria que especifica si el paciente ha recibido un transplante o no. Esta variable tomará el valor 1 cuando el paciente ha recibido un transplante y 0 en caso contrario. Se trata de una variable dependiente del tiempo.
- *id*: identificador numérico de cada uno de los pacientes.

Una vez presentadas cada una de las variables del conjunto de datos, se extraen los datos directamente desde R con la siguiente orden:

```
data("heart", package = "survival")
```

En la siguiente tabla se adjunta una muestra del *data.frame* en estudio:

start	*stop*	*event*	*age*	*year*	*surgery*	*transplant*	*id*
0	46	0	0.9253936	2.507871	0	0	36
46	100	1	0.9253936	2.507871	0	1	36
0	78	0	-6.617385	3.994524	1	0	59
78	916	0	-6.617385	3.994524	1	1	59
0	6	0	4.892539	5.284052	0	0	81
6	445	0	4.892539	5.284052	0	1	81
0	21	0	-0.9007529	3.340178	0	0	47
21	72	1	-0.9007529	3.340178	0	1	47
0	60	0	-1.746749	5.470226	0	0	87
60	110	1	-1.746749	5.470226	0	1	87

Tabla 4.19: Muestra de datos del conjunto *heart* contenido en el paquete *survival*.

Tras la comprensión de la estructura de los datos, se procede a la construcción de un modelo de Cox para la variable tiempo de supervivencia expresada mediante intervalos de tiempo de la forma *(tstart, tstop]* frente al resto de variables disponibles en el conjunto de datos bajo estudio. Entre estas variables disponibles, se encuentran dos variables no dependientes del tiempo (*age* y *year*) y dos variables dependientes del tiempo o variantes en el tiempo (*surgery* y *transplant*). Todas estas variables actuarán como variables explicativas o predictoras en el modelo de regresión a construir:

```
mod_cox_dep_t <- coxph(Surv(start, stop, event) ~ transplant + age +
  year + surgery, data = heart)
```

A continuación, se muestra la información que posee el objeto *coxph* creado con la orden anterior:

```
summary(mod_cox_dep_t)
```

```
## Call:
## coxph(formula = Surv(start, stop, event) ~ transplant + age +
##     year + surgery, data = heart)
##
##   n= 172, number of events= 75
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## transplant1 -0.01025   0.98980  0.31375 -0.033   0.9739
## age          0.02717   1.02754  0.01371  1.981   0.0476 *
## year        -0.14635   0.86386  0.07047 -2.077   0.0378 *
## surgery     -0.63721   0.52877  0.36723 -1.735   0.0827 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## transplant1    0.9898     1.0103    0.5352    1.8307
## age            1.0275     0.9732    1.0003    1.0555
## year           0.8639     1.1576    0.7524    0.9918
## surgery        0.5288     1.8912    0.2574    1.0860
##
## Concordance= 0.636  (se = 0.033 )
## Likelihood ratio test= 15.11  on 4 df,   p=0.004
## Wald test            = 14.49  on 4 df,   p=0.006
## Score (logrank) test = 15.03  on 4 df,   p=0.005
```

La información que se puede consultar en la salida obtenida es la misma que se resaltó en el Apartado 4.3.1 del presente trabajo. Entre ella, es destacable el número de observaciones utilizadas a la hora de construir el modelo, en este caso, 172; o el número de observaciones para las que se ha producido el evento de interés (muerte del paciente que se encuentra a la espera de recibir un transplante de corazón), 75. El resto de información relevante se analizará en las próximas secciones.

A continuación, se procede al estudio detallado del modelo con la estimación de parámetros, la significación de estos y la evaluación de su calidad predictiva. Recuérdese que en el modelo de Cox con covariantes dependientes del tiempo no hace falta estudiar si se verifica la hipótesis de riesgos proporcionales pues, en general, el modelo de Cox extendido para este tipo de variables no satisface dicha hipótesis (véase el Apartado 3.4).

4.3.2.1. Estimación de parámetros

El estudio de los parámetros o coeficientes estimados para cada una de las variables involucradas en el modelo se realizará con la siguiente orden:

```
mod_cox_dep_t$coefficients
```

Variable	*Estimación*
transplant1	-0.01025077
age	0.02716664
year	-0.14634635
surgery	-0.63720989

Tabla 4.20: Estimación de los coeficientes asociados a cada una de las variables explicativas del modelo creado con covariantes dependientes del tiempo.

Para una interpretación de los coeficientes anteriores y el efecto que tienen sobre el tiempo de supervivencia modelizado en intervalos de tiempo, es necesario estudiar la exponencial de dichas estimaciones. A continuación, se adjunta una tabla en la que se pueden consultar dichas exponenciales para cada uno de los parámetros estimados junto a su intervalo de confianza al 95 %:

```
summary(mod_cox_dep_t)$conf.int
```

Variable	*exp(coef)*	*lower .95*	*upper .95*
transplant1	0.9898016	0.5351550	1.8306980
age	1.0275390	0.7524197	0.9918021
year	0.8638585	0.8066522	1.5376930
surgery	0.5287657	0.2574423	1.0860419

Tabla 4.21: Estimación de la exponencial de los coeficientes asociados a cada una de las variables explicativas del modelo creado con covariantes dependientes del tiempo junto a su intervalo de confianza al 95 %.

Para evidenciar el efecto que tienen las variables involucradas en el modelo sobre la función de riesgo de muerte de los pacientes que se encuentran a la espera de un transplante cardíaco, se analizarán dos variables:

- Variable *age*: se ha obtenido que $\hat{\beta}_2 = 0.02716664$ y $e^{\hat{\beta}_2} = 1.0275390$, de modo que, manteniendo constantes el resto de variables involucradas en el modelo de Cox creado, el aumento en un año de la edad de un paciente supone que la función riesgo de muerte de este aumente de manera aproximada en un 2.75 %, ya que $(1.0275390 - 1) \cdot 100\,\% \approx 2.75\,\%$.
- Variable *year*: se tiene que $\hat{\beta}_3 = -0.14634635$ y $e^{\hat{\beta}_3} = 0.8638585$, por lo que, si se mantienen constantes el resto de variables incluidas en el modelo de Cox construido, el entrar en el programa un año más tarde le supone a un paciente que la función riesgo de muerte disminuya en un factor de 0.8638585, que en promedio se corresponde con un 13.6 % aproximadamente, dado que $(1 - 0.8638585) \cdot 100\,\% \approx 13.61\,\%$.

Nótese que esta interpretación detallada de los coeficientes sobre el efecto que tienen en la función de riesgo de muerte asociada a los pacientes debe realizarse por un experto en la materia. Además, se puede estudiar el intervalo de confianza asociado a cada uno de los parámetros estimados de la siguiente manera:

```
confint(mod_cox_dep_t)
```

Variable	*coef*	*2.5 %*	*97.5 %*
transplant1	-0.01025077	-0.6251988771	0.604697332
age	0.02716664	0.0002874691	0.054045813
year	-0.14634635	-0.2844610476	-0.008231644
surgery	-0.63720989	-1.3569596167	0.082539837

Tabla 4.22: Estimación de los coeficientes asociados a cada una de las variables explica-

tivas del modelo creado con covariantes dependientes del tiempo junto a su intervalo de confianza al 95 %.

En el estudio del modelo, es posible consultar también el número de iteraciones, en este caso 4, que ha tenido que realizar el algoritmo para que el método converja:

```
mod_cox_dep_t$iter
```

```
## [1] 4
```

Asimismo, puede consultarse el valor de la log-verosimilitud del modelo junto al mismo valor bajo la hipótesis nula $H_0 : \underline{\beta} = \underline{0}$:

```
mod_cox_dep_t$loglik
```

```
## [1] -298.1214 -290.5656
```

4.3.2.2. Contrastes de hipótesis

Tras la estimación de los parámetros asociados a las variables predictoras consideradas, es necesario estudiar si estos son significativos o no para el modelo de regresión construido. Es decir, se deben resolver los siguientes contrastes de hipótesis para cada una de las variables explicativas consideradas en el modelo creado; i.e.:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \end{cases}$$

para todo $j = 1, \ldots, 4$ donde, por simplicidad, los índices $j = 1$ y $j = 4$ hacen referencia a las variables dependientes del tiempo. Para alcanzar dicho objetivo, se analiza la columna *Pr(>|z|)* de la siguiente tabla que contiene los p-valores asociados a cada una de las estimaciones obtenidas mediante el estadístico de Wald que se adjunta también en la columna z:

```
summary(mod_cox_dep_t)$coef[, c(4, 5)]
```

Variable	z	*Pr(>|z|)*
transplant1	-0.03267128	0.97393672
age	1.98092553	0.04759963 *
year	-2.07677794	0.03782206 *
surgery	-1.73519821	0.08270570 .

Tabla 4.23: Contrastes de hipótesis individuales asociados al modelo de Cox creado con covariantes dependientes del tiempo.

Consultando los p-valores obtenidos, se concluye que las variables significativas del modelo creado a un nivel de significación del 5 % son *age* y *year*, ambas variables no dependientes del tiempo. En contraposición, las variables dependientes del tiempo consideradas en el modelo (*transplant* y *surgery*) son no significativas. Sin embargo, en el caso de que se considerase el estudio a un nivel de significación del 10 %, la variable *surgery* sería también significativa en el modelo.

Una vez analizados los contrastes de hipótesis individuales, debe estudiarse también lo que ocurre con el contraste de hipótesis conjunto de todas las variables explicativas usadas en el modelo de Cox. Este se enuncia a continuación:

$$\begin{cases} H_0 : \underline{\beta} = (\beta_{10}, \ldots, \beta_{40})' = \underline{0} \\ H_1 : \beta_j \neq \beta_{j_0} \text{ para algún } j = 1, \ldots, 4, \end{cases}$$

donde al igual que se enunció anteriormente, los índices $j = 1$ y $j = 4$ hacen referencia a las variables dependientes del tiempo.

Dicho contraste de hipótesis se puede resolver a través de tres test diferentes (test de razón de verosimilitud, test de Wald y test score), los cuales son asintóticamente equivalentes. En la tabla siguiente se pueden examinar los estadísticos χ^2 con sus grados de libertad y p-valores asociados a cada uno de los test expuestos:

```
t(data.frame(
  "Test de razón de verosimilitud" = summary(mod_cox_dep_t)$logtest,
  "Test de Wald" = summary(mod_cox_dep_t)$waldtest,
  "Test score" = summary(mod_cox_dep_t)$sctest
))
```

	test	*df*	*pvalue*
Test de razón de verosimilitud	15.11148	4	0.004475500
Test de Wald	14.49000	4	0.005876870
Test score	15.03420	4	0.004630811

Tabla 4.24: Contraste de hipótesis conjunto asociado al modelo de Cox creado con covariantes dependientes del tiempo.

Como todos los p-valores obtenidos son significativos, se concluye que el modelo creado permite explicar la variable tiempo hasta la muerte del paciente.

Para finalizar el estudio de los parámetros y los contrastes de hipótesis asociados a ellos, se adjunta una gráfica que reúne toda la información relevante para el modelo. En dicha gráfica se observarán las cuatro variables predictoras consideradas junto a la exponencial de sus coeficientes estimados. Asimismo, para cada uno de los coeficientes se mostrará un gráfico de la estimación realizada junto a su intervalo de confianza y p-valor asociado. Una explicación más detallada del contenido de la gráfica se puede consultar en la aplicación expuesta en la página 58 para la Figura 4.6. Finalmente, la orden necesaria para obtener la gráfica en cuestión se enuncia a continuación:

```
ggforest(mod_cox_dep_t, main = "Resumen del modelo")
```

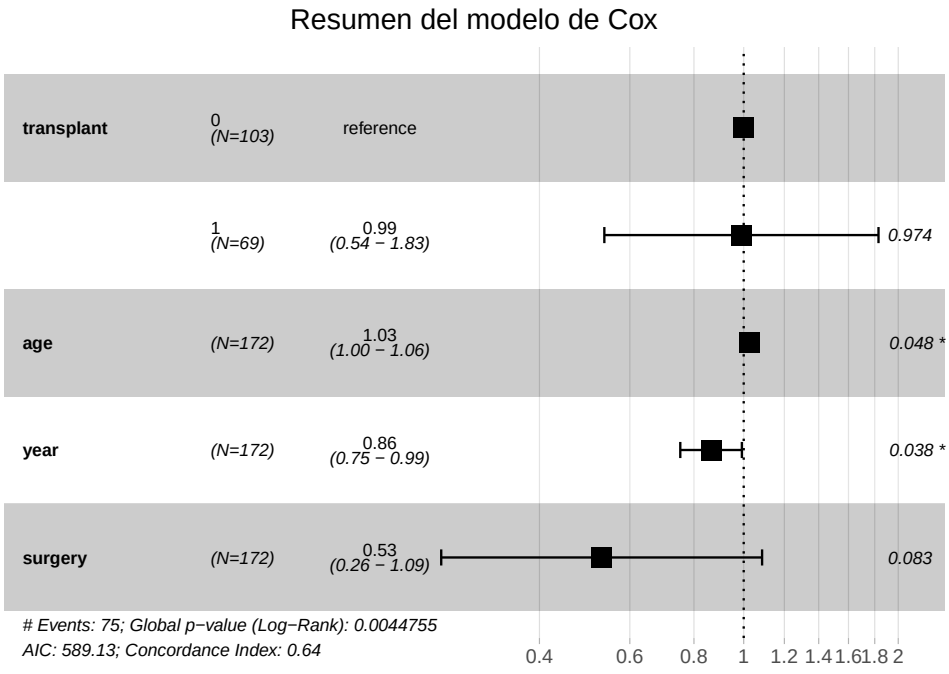


Figura 4.12: Resumen gráfico del análisis del modelo de Cox con covariantes dependientes del tiempo realizado con la función *ggforest()* del paquete *survminer*.

4.3.2.3. Evaluación del modelo

Tal y como se indicó en el Apartado 3.7 de la parte teórica de la evaluación del modelo de Cox con covariantes dependientes del tiempo, la calidad predictiva del modelo se puede estudiar a través de medidas análogas al coeficiente de determinación R^2. Estas medidas incluyen el coeficiente sugerido por Nagelkerke (1991), el propuesto por O'Quigley et al. (2005) y el sugerido por Royston (2006). Al igual que se especificó en Apartado 4.3.1.4 de la aplicación anterior del modelo de Cox que incluía solo covariantes independientes del tiempo, dichos coeficientes se pueden calcular y mostrar de la siguiente forma:

```
rsq(mod_cox_dep_t, sigD = NULL)
```

```
## $cod
## [1] 0.08410856
##
## $mer
## [1] 0.1824853
##
## $mev
## [1] 0.1194867
```

Dado que los coeficientes obtenidos son muy cercanos al 0, se deduce que la calidad del modelo creado con covariantes dependientes del tiempo no es excesivamente buena para la predicción del tiempo de supervivencia. Sin embargo, debe tenerse en cuenta que el modelo obtenido es válido para determinar aquellas variables que influyen en dicho tiempo y el sentido de dicha influencia.

A pesar de que en un modelo de Cox con covariantes dependientes del tiempo no hace falta estudiar si las variables involucradas en el modelo cumplen la hipótesis de riesgos proporcionales o no, a continuación, se presenta un pequeño estudio para analizar qué es lo que ocurre con dicha hipótesis en el modelo bajo estudio.

Al igual que se mostró en la diagnosis realizada sobre el modelo de Cox con covariantes independientes del tiempo del apartado anterior, el estudio de si la hipótesis de riesgos proporcionales se viola o no, se puede hacer de manera gráfica y analítica mediante el test de bondad de ajuste (GOF).

En la siguiente tabla, se pueden analizar los resultados obtenidos tras resolver los diferentes contrastes de bondad de ajuste asociados a cada una de las variables predictoras del modelo de Cox creado:

```
cox.zph(mod_cox_dep_t, transform = "rank")
```

Variable	*chisq*	*df*	*p*
transplant	0.25244	1	0.62
age	0.83072	1	0.36
year	1.05187	1	0.31
surgery	0.00385	1	0.95
GLOBAL	2.68433	4	0.61

Tabla 4.25: Estudio de la hipótesis de riesgos proporcionales de manera individual y global para el modelo de Cox creado con covariantes dependientes del tiempo.

A la vista de los p-valores obtenidos para cada contraste realizado, se concluye que no existen evidencias significativas en contra de la hipótesis nula de dichos contrastes. Como consecuencia de ello, todas las variables predictoras del modelo creado verifican la hipótesis de riesgos proporcionales a nivel individual. Asimismo, dicha hipótesis se cumple de manera conjunta para todas las variables involucradas en el modelo. Nótese que en un modelo de Cox con covariantes dependientes del tiempo, el hecho de que se cumpla la hipótesis de riesgos proporcionales para cada una de las variables explicativas no es usual. Sin embargo, en el caso bajo estudio, esta condición se cumple ya que, tal y como se ha comprobado en el Apartado 4.3.2.2, las variables dependientes del tiempo son no significativas y pueden considerarse nulos sus coeficientes estimados.

Para el estudio gráfico de las hipótesis de riesgos proporcionales, es necesario calcular los residuos de Schoenfeld asociados a cada uno de los individuos del conjunto de datos bajo estudio y a cada variable explicativa considerada en el modelo. A continuación, se muestra el cálculo de estos residuos escalados:

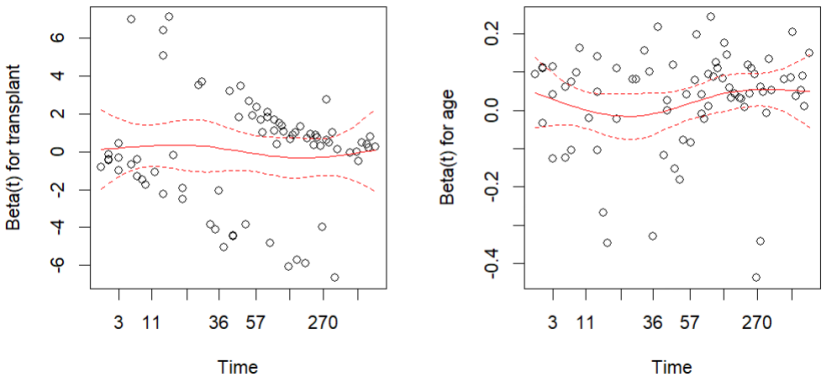
```
cox.zph(mod_cox_dep_t, transform = "rank")$y
```

time	*transplant*	*age*	*year*	*surgery*
1	-0.8012696	0.09340004	-1.1193608	9.347928
2	-0.3913630	0.11085396	0.7490845	-2.239041
2	-0.4647442	0.10960172	0.4462945	-1.915566
2	-0.1436648	-0.03385701	-0.4402613	8.687208
3	-0.3362103	0.04168269	0.2541290	-1.707941
3	0.4353129	-0.12542568	-0.1590904	-1.309205
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
733	-0.5020444	0.204263442	-0.19367658	-2.618850
852	0.4910750	0.035798724	-0.13991515	-3.136782
980	0.3940582	0.051646341	0.09952175	6.115883
996	0.2148273	0.090662022	0.09884397	6.524236
1032	0.7810621	0.009864047	-0.25340874	-2.322435
1387	0.2522059	0.149259359	-0.30396670	-1.402432

Tabla 4.26: Residuos escalados de Schoenfeld asociados al modelo de Cox creado con covariantes dependientes del tiempo.

Una vez obtenidos los residuos de Schoenfeld, se procede a realizar unas gráficas de estos para su posterior análisis:

```
par(mfrow = c(1, 2))
plot(cox.zph(mod_cox_dep_t, transform = "rank"), col = "red")
par(mfrow = c(1, 1))
```



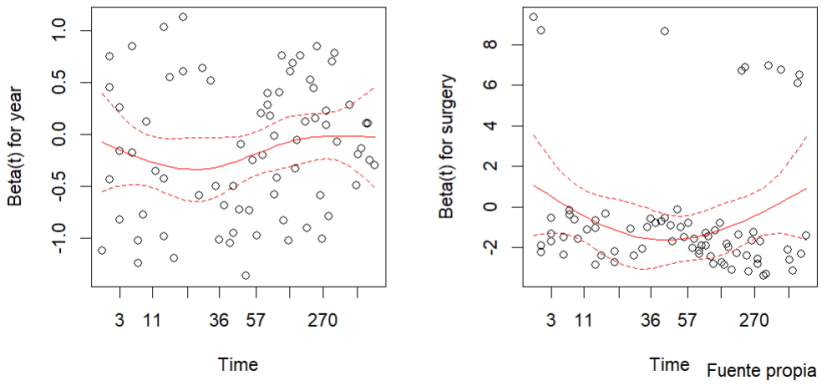


Figura 4.13: Gráficas de los residuos escalados de Schoenfeld con la orden *plot(cox.zph())* para el modelo de Cox creado con covariantes dependientes del tiempo.

Observando las gráficas obtenidas, no se tiene muy claro si las variables usadas en el ajuste del modelo cumplen o no la hipótesis de riesgos proporcionales pues no se ve claramente si existen patrones en los residuos. Ante esta dificultad, se presenta otra forma de obtener las gráficas de los residuos escalados de Schoenfeld mucho más interpretable. Esta alternativa hace uso de la función *ggcoxzph()* y se aplica de la siguiente forma:

```
ggcoxzph(cox.zph(mod_cox_dep_t),
  font.main = 10, font.submain = 8, font.x = 10,
  font.y = 10, font.tickslab = 10, font.legend = 8
)
```

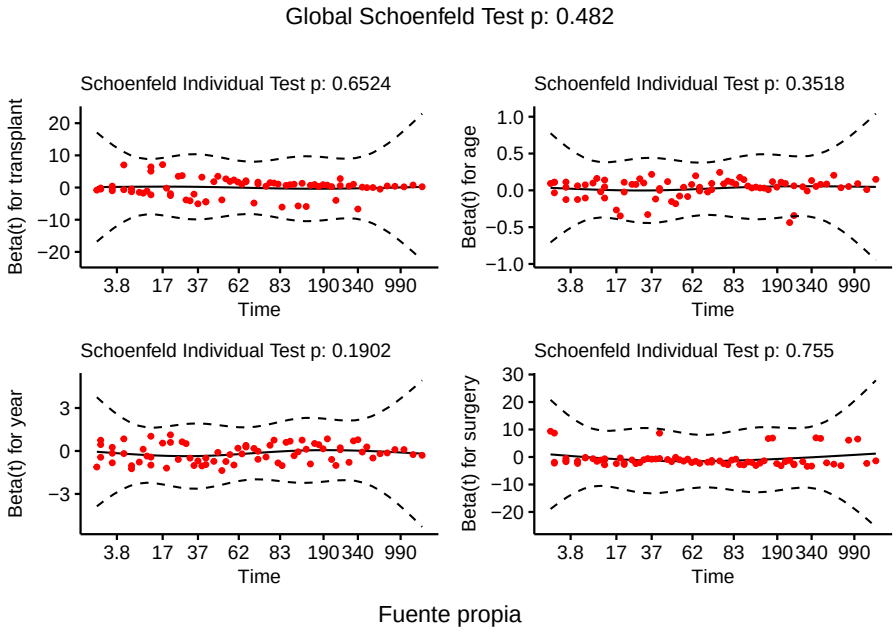


Figura 4.14: Estudio gráfico y analítico de los residuos de Schoenfeld con la orden *ggcoxzph()* de la librería *survminer*.

Tal y como se observa en las gráficas resultantes, esta función tiene la peculiaridad de que además de mostrar la gráfica de los residuos, la salida adjunta los p-valores asociados a los test de hipótesis de bondad de ajuste propuestos por Schoenfeld que analizan si cada una de las variables violan o no la hipótesis de riesgos proporcionales. De esta forma, se obtiene un resumen completo del estudio de las hipótesis de riesgos proporcionales de manera gráfica y analítica. Gráficamente, las variables violarán la hipótesis de riesgos proporcionales cuando los residuos de Schoenfeld no se encuentren entre las curvas discontinuas y además, no se distribuyan alrededor de la curva continua; y, analíticamente, las variables violarán la hipótesis de riesgos proporcionales cuando los p-valores asociados a estas, sean menores que el nivel de significación prefijado.

Una vez expuesto cómo se interpreta la salida, se concluye de forma clara que el modelo de Cox obtenido verifica las hipótesis de riesgos proporcionales tanto gráficamente como analíticamente. De manera individual, todos los p-valores obtenidos son mayores que 0.05 y todos los residuos se distribuyen alrededor de la recta continua, encontrándose además, entre las curvas discontinuas, por lo que se concluye que todas las variables explicativas del modelo verifican la hipótesis de riesgos proporcionales. Por su parte, de manera global, el p-valor asociado al contraste correspondiente es mayor que el nivel de significación prefijado, por lo que no existen evidencias significativas en contra de la hipótesis de riesgos proporcionales.

4.3.2.4. Selección de variables del modelo

En esta sección, se llevará a cabo un estudio de selección de variables de un modelo de Cox que contiene tanto covariantes independientes como dependientes del tiempo. Hasta ahora se ha estado usando el conjunto de datos *heart* de la librería *survival* para la construcción de un modelo de Cox con covariantes dependientes del tiempo pero, dado que este posee pocas variables, se ha estimado más conveniente utilizar el conjunto de datos *pbc_unif* anteriormente trabajado para mostrar la utilidad y manejo de la función *tmerge()*, ya que contiene muchas más variables.

Antes de comenzar a usar las funciones disponibles en la librería *glmnet* para el estudio de selección de variables, es necesario hacer una limpieza de los datos y extraer el subconjunto de datos que no posee observaciones perdidas:

```
pbc_unif_na <- na.omit(pbc_unif)
```

Asimismo, tal y como se especificó en el Apartado 4.3.1.5, es imprescindible que los datos sean de tipo numérico y que existan dos conjuntos de datos: uno con las variables explicativas que se quieran considerar, y otro con las variables que contienen la información del tiempo y censura o estado (esta última variable solo admite los valores 0 y 1). En el caso de modelos que contengan covariantes dependientes del tiempo existe una salvedad a la hora de definir el conjunto de datos que posee la información relativa al tiempo de supervivencia y censura, ya que el tiempo no es puntual sino un intervalo. Esta salvedad puede solucionarse definiendo el conjunto como un objeto de tipo *Surv* con la estructura *Surv(tstart, tstop, status)*. Además, el conjunto que contiene a las variables predictoras debe ser una matriz; y en el caso de estudio, dado que la variable *status* toma tres posibles valores, es necesario cambiar el valor 2 por otro valor que será el 1:

- Preparación del conjunto de datos que contiene las variables explicativas que se quieren considerar para la construcción y ajuste del modelo:

```
x_pbc <- as.matrix(pbc_unif_na[, c(
  "trt", "age", "ascites", "hepato",
  "edema", "bili", "chol", "albumin",
  "copper", "alk.phos", "ast", "trig",
  "platelet", "protime"
)])
```

- Creación de una nueva variable *status* que solo toma los valores 0 y 1:

```
pbc_unif_na$status1 <- ifelse(pbc_unif_na$status == 2, 1, 0)
```

- Construcción del objeto *Surv* que contiene la información referente al tiempo de supervivencia y censura:

```
y_pbc <- Surv(
  pbc_unif_na$tstart,
  pbc_unif_na$tstop,
  pbc_unif_na$status1
)
```

Una vez preparados los datos, se procede a la construcción del ajuste usando validación cruzada y el método Lasso con la función *cv.glmnet()* para así encontrar los valores óptimos del parámetro λ. Dicha función tomará los argumentos *family = "cox"*, *type.measure = "deviance"* y *nlambda = 100* descritos en el uso de esta función para el modelo de Cox con variables independientes del tiempo (Apartado 4.3.1.5):

```
cvfit1 <- cv.glmnet(
  x_pbc, y_pbc,
  family = "cox",
  type.measure = "deviance",
  nlambda = 100
)
```

Tras realizar el ajuste, es de utilidad mostrar una gráfica en la que se puedan consultar los valores óptimos $\lambda_{\text{mín}}$ y λ_{1se} junto con el número de variables consideradas en cada ajuste y la evolución del mismo. En ella se observará que para ambos parámetros óptimos se ha producido una selección de variables, consecuencia directa con el hecho de que se está realizando un ajuste usando el método Lasso:

```
plot(cvfit1)
```

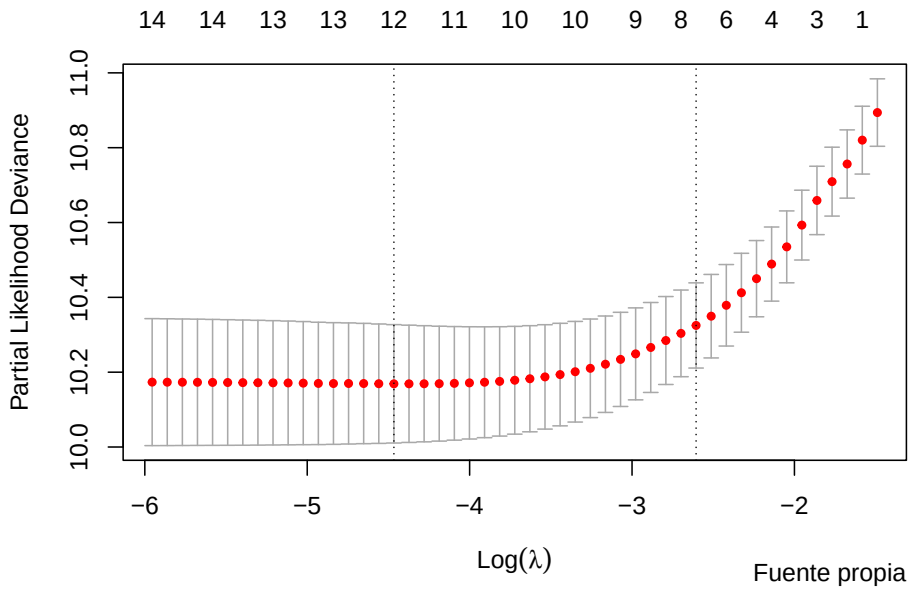


Figura 4.15: Evolución del valor de $log(\lambda)$ en función de la verosimilitud parcial obtenida con la función *cv.glmnet()* para el modelo de Cox con covariantes dependientes del tiempo.

Una vez mostrada la gráfica, se extraen los valores óptimos de λ:

```
data.frame(cvfit1$lambda.min, cvfit1$lambda.1se)
```

Parámetro	*Estimación*
$\lambda_{\text{mín}}$	0.01148991
λ_{1se}	0.07385799

Tabla 4.27: Ajuste de los parámetros en la selección de variables en el modelo de Cox con covariantes dependientes del tiempo.

A continuación, para cada uno de los λ's óptimos se construye el modelo correspondiente con la orden *glmnet()*:

1. Modelo usando $\lambda_{\text{mín}}$:

```
fit_dep <- glmnet(x_pbc, y_pbc,
  family = "cox",
  lambda = cvfit1$lambda.min
)
```

Después de haber definido el modelo, se presenta una tabla en la que se pueden consultar los valores de los coeficientes estimados:

```
fit_dep$beta
```

Variable	*s0*
trt	-0.0264604168
age	0.0220063613
ascites	.
hepato	0.2928600539
edema	0.1457894297
bili	0.0501542642
chol	.
albumin	-0.6142637525
copper	0.0017648608
alk.phos	0.0001224687
ast	0.0017077558
trig	0.0015381940
platelet	-0.0002798852
protime	0.0667553416

Tabla 4.28: Estimación de los coeficientes asociados a cada una de las variables explicativas usando $\lambda_{\text{mín}}$ en el método Lasso para un modelo de Cox con covariantes dependientes del tiempo.

A la vista de los coeficientes estimados, se ha obtenido un modelo en el que dos de las variables explicativas consideradas en el modelo, *ascites* y *chol*, son nulas.

2. Modelo usando λ_{1se}:

```
fit_dep1 <- glmnet(x_pbc, y_pbc,
  family = "cox",
  lambda = cvfit1$lambda.1se
)
```

Tras la definición del modelo, se muestran los coeficientes estimados para cada una de las variables predictoras involucradas en el modelo:

```
fit_dep1$beta
```

Variable	*s0*
trt	.
age	6.716088e-03
ascites	.
hepato	1.340055e-01
edema	.
bili	5.179032e-02
chol	.
albumin	-3.407352e-01
copper	1.611880e-03
alk.phos	4.945505e-06
ast	.
trig	.
platelet	.
protime	6.043381e-02

Tabla 4.29: Estimación de los coeficientes asociados a cada una de las variables explicativas usando λ_{1se} en el método Lasso para un modelo de Cox con covariantes dependientes del tiempo.

Tal y como se puede consultar en la tabla, se ha conseguido un nuevo modelo con siete variables nulas. De este modo, el tiempo hasta la muerte del paciente puede ser modelizada por otras siete variables, consiguiendo así un modelo mucho más sencillo que el creado con $\lambda_{\text{mín}}$. Cabe destacar también que las estimaciones no nulas de las variables explicativas son prácticamente 0.

Nótese que en ocasiones, pueden darse problemas de convergencia a la hora de buscar los parámetros óptimos con las funciones *cv.glmnet()* y *glmnet()* expuestos en este apartado.

Para finalizar el estudio de selección de variables en un modelo de Cox, es útil mostrar una gráfica en la que se puede consultar la evolución de los valores de los coeficientes de las variables explicativas consideradas frente a la norma L1 a medida que el valor del parámetro λ varía. Asimismo, en el eje superior de la gráfica que se adjunta a continuación, se puede ir consultando el número de coeficientes no nulos:

```
plot(glmnet(x_pbc, y_pbc, family = "cox"))
```

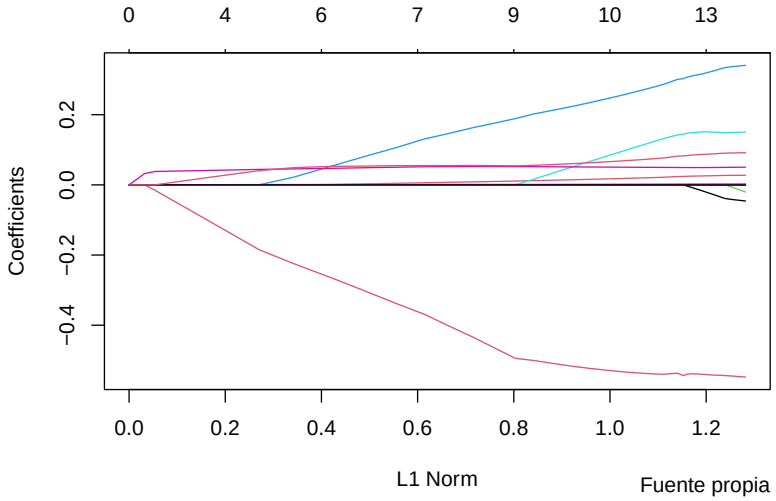


Figura 4.16: Evolución del valor de los coeficientes de las variables predictoras frente a la norma L1 en el método Lasso para un modelo de Cox con covariantes dependientes del tiempo.

4.3.2.5. Otras consideraciones del modelo

Finalmente, se adjuntan órdenes y salidas que pueden ser de utilidad a la hora de analizar de manera exhaustiva un modelo de Cox. Para ello, el modelo con covariantes dependientes del tiempo que se usará es el que se ha programado con el nombre *mod_cox_dep_t*.

Para empezar, se muestra la estimación de la matriz de varianzas y covarianzas de las variables predictoras consideradas en el modelo para así evaluar la relación existente entre ellas:

```
mod_cox_dep_t$var
```

```
##                [,1]          [,2]          [,3]          [,4]
## [1,]  0.0984420735 -9.626259e-04  0.0007419972 -3.719111e-03
## [2,] -0.0009626259  1.880770e-04  0.0001080974 -6.702935e-05
## [3,]  0.0007419972  1.080974e-04  0.0049657361 -5.282355e-03
## [4,] -0.0037191108 -6.702935e-05 -0.0052823549  1.348549e-01
```

Cuando se construye un modelo de regresión de Cox es posible estudiar la distribución de la función de supervivencia $S(t)$ asociada. Para ello, la estimación se hará usando el método de Kaplan-Meier con la siguiente orden:

```
func_superv_dep <- survfit(mod_cox_dep_t, type = "kaplan-meier")
```

Otra forma equivalente de obtener dicha estimación es mediante la siguiente orden:

```
survfit(mod_cox_dep_t, stype = 1)
```

Tras el cálculo de su estimación, se presenta un resumen en el que se puede consultar para cada tiempo observado el número de individuos en riesgo (*n.risk*), el número de individuos que presentan el evento en dicho instante (*n.event*), la estimación de $S(t)$ (*survival*) junto con su intervalo de confianza al 95 % (*lower 95 % CI* y *upper 95 % CI*) y la desviación estándar de la estimación realizada (*std.err*):

```
summary(func_superv_dep)
```

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower 95 % CI*	*upper 95 % CI*
1	103	1	0.990	0.00956	0.9718	1.000
2	102	3	0.962	0.01895	0.9253	1.000
3	99	3	0.933	0.02479	0.8856	0.983
5	96	2	0.914	0.02797	0.8605	0.970
6	94	2	0.894	0.03074	0.8361	0.957
8	92	1	0.885	0.03205	0.8239	0.950
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
733	16	1	0.236	0.08937	0.1126	0.496
852	14	1	0.217	0.08878	0.0974	0.484
980	11	1	0.196	0.08791	0.0812	0.472
996	10	1	0.176	0.08608	0.0671	0.459
1032	9	1	0.155	0.08296	0.0545	0.442
1387	6	1	0.130	0.07748	0.0405	0.418

Tabla 4.30: Estimación de la función de supervivencia a partir del modelo de Cox creado con covariantes dependientes del tiempo.

Por último, para completar el estudio de la función de supervivencia, se hace una gráfica de su estimación para todo el conjunto de datos involucrado en el modelo de Cox creado con covariantes dependientes del tiempo:

```
ggsurvplot(survfit(mod_cox_dep_t),
  data = uissurv, conf.int = T,
  xlab = "t (días)", ylab = "S(t) estimada",
  legend.title = "Estimación",
  legend.labs = "Kaplan-Meier",
  ggtheme = theme_grey()
)
```

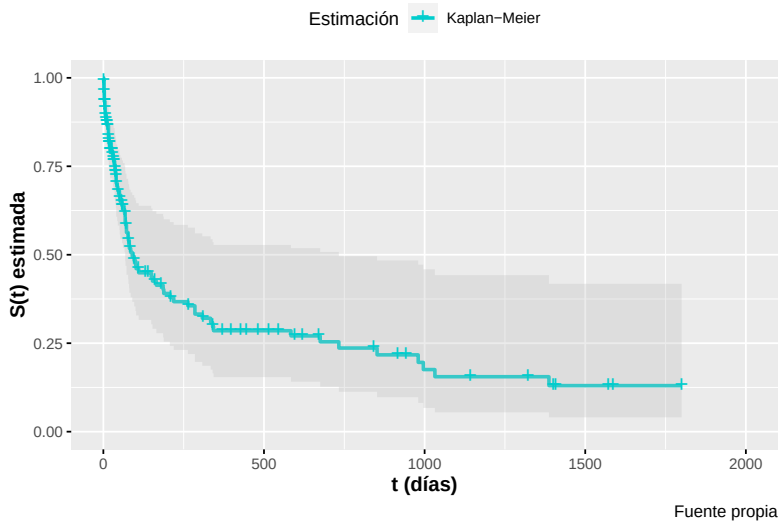


Figura 4.17: Estimación de la función de supervivencia para el modelo creado y ajustado con covariantes dependientes del tiempo.

Conclusión

El análisis de supervivencia es una técnica muy útil cuando el objetivo del problema o análisis a realizar es el estudio y modelización de la variable tiempo que transcurre hasta que se produce un determinado suceso de interés. Se trata también de una rama de la estadística en continua evolución y estudio, a la que todavía le queda mucho por avanzar tal y como se puede comprobar en los artículos recientes que se han utilizado como bibliografía a la hora de la realización de este trabajo. Asimismo, el análisis de supervivencia destaca por las aplicaciones que tiene hoy en día sobre todo en disciplinas como la medicina o la ingeniería.

A lo largo del trabajo, se ha realizado un estudio detallado sobre el modelo de Cox con covariantes dependientes e independientes del tiempo. Sin embargo, dada la magnitud de la materia, no se ha podido profundizar en algunos aspectos citados en el presente Trabajo Fin de Máster. A continuación, se enumeran diferentes conceptos y temas que en un posterior estudio se podrían profundizar:

- Propiedades asintóticas tanto del modelo de Cox con covariantes dependientes del tiempo como dicho modelo con covariantes independientes del tiempo.
- Estudio de aplicaciones del modelo de Cox con covariantes dependientes del tiempo en el ámbito científico, médico...
- Estimación de parámetros para el modelo de Cox con covariantes dependientes del tiempo y tipos de censuras diferentes a la censura a la derecha. Nótese que para la censura a la izquierda bastaría con hacer una pequeña transformación. Tal y como determinan Goel y Klein (1992), la obtención de datos censurados a la derecha partiendo de datos censurados a la izquierda se consigue con el producto de cada uno de los datos por -1. De esta forma, cualquier técnica o estudio realizado para los datos censurados a la derecha se pueden extrapolar a los datos censurados a la izquierda.
- Estimación de parámetros del modelo de Cox con covariantes dependientes del tiempo cuando existen tiempos de fallos coincidentes o empates.
- A la hora de resolver la estimación de los parámetros de cada uno de los modelos de Cox expuestos a lo largo del trabajo, se necesita resolver un sistema de ecuaciones no lineales cuya expresión no se detalla en el presente trabajo. Se podría ampliar con el análisis de los procedimientos iterados de aproximación a la solución de los problemas de optimización

Bibliografía

O. Aalen, O. Borgan y H. Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

M.P. Allen. *Understanding regression analysis*. Springer Science & Business Media, 1997. doi: https://doi.org/dnkkn9.

A. Allignol y A. Latouche. *CRAN Task View: Survival Analysis*. R Foundation for Statistical Computing, Version 2022-03-07. Disponible en: https://cran.r-project.org/web/views/Survival.html.

E.E. Alvarez y J. Ferrario. Revisión de estimación robusta en modelos semiparamétricos de supervivencia. *Estadistica*, 64, 2012.

C.A. Bellera, G. MacGrogan, M. Debled, C.T. De Lara, V. Brouste y S. Mathoulin-Pélissier. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC medical research methodology*, 10(1):1–12, 2010.

R.A. Berk, P.H. Rossi y K.J. Lenihan. *Money, Work, and Crime: Experimental Evidence*. New York: Academic Press, 1980.

P. Biecek, S. Fabian, A. Kassambara y M. Kosinski. *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.9, 2021. Disponible en: https://CRAN.R-project.org/package=survminer.

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30:89–99, 1974.

R.H. Chan y J. Ma. A multiplicative iterative algorithm for box-constrained penalized likelihood image restoration. *IEEE transactions on image processing*, 21(7):3168–3181, 2012.

D. Cox y D. Oakes. *Analysis of survival data*, volume 21. CRC press, 1984.

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972. doi: https://doi.org/gf2fbn.

D.R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

J. Crowley y M. Hu. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72(357):27–36, 1977.

C. Dardis. *survMisc: Miscellaneous Functions for Survival Data*. R package version 0.5.6, 2022. Disponible en: https://CRAN.R-project.org/package=survMisc.

E. de las Heras. Comparación de curvas de supervivencia con datos censurados. Trabajo Fin de Grado, Universidad de Sevilla. Disponible en https://hdl.handle.net/11441/134450, 2021.

P. Deuflhard. *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*, volume 35. Springer Science & Business Media, 2011. doi: https://doi.org/dv8qcx.

B. Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

L. Fahrmeir, G. Tutz, W. Hennevogl y E. Salem. *Multivariate statistical modelling based on generalized linear models*, volume 425. Springer, 1994.

L.D. Fisher y D.Y. Lin. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157, 1999.

J. Fox y S. Weisberg. *An R companion to applied regression*. Sage publications, 2018.

J. Fox y S. Weisberg. Cox proportional-hazards regression for survival data. *An R companion to applied regression*, 2018, 2023.

J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, J. Qian y J. Yang. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 4.1.8., 2023. Disponible en: https://CRAN.R-project.org/package=glmnet.

P. Goel y J.P. Klein. *Survival Analysis: State of the Art*. Springer Science & Business Media, First edition, 1992.

A. González. Selección de variables: Una revisión de métodos existentes. Trabajo Fin de Máster, Universidad de los Andes. Disponible en http://eio.usc.es/pub/mte/index.php/es/trabajos-fin-de-master/finalizados, 2015.

I. Guyon y A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

G. Gómez y C. Cadarso-Suárez. El modelo de riesgos proporcionales de Cox y sus extensiones. Impacto en Estadística y Biomedicina. *La Gaceta de la RSME*, 20(3):513–538, 2017.

M.A. Gómez. Karl Pearson, el creador de la estadística matemática. *Historia de la Probabilidad y la estadística (IV)*, páginas 351–356, 2009.

F.E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer, 2001.

F.E. Harrell y K.L. Lee. Verifying assumptions of the Cox proportional hazards model. In *Proceedings of the eleventh annual SAS Users group international conference*, páginas 823–828. SAS Institute Inc, Cary, NC, 1986.

F.E. Harrell, K.L. Lee y D.B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

E. Harrison, T. Drake y R. Ots. *finalfit: Quickly Create Elegant Regression Results Tables and Plots when Modelling*. R package version 1.0.6, 2023. Disponible en: https://CRAN.R-project.org/package=finalfit.

P. Hougaard. *Analysis of multivariate survival data*, volume 564. Springer, 2000.

K. Ito y K. Kunisch. A variational approach to sparsity optimization based on Lagrange multiplier theory. *Inverse problems*, 30(1):015001, 2013.

C.M. Jareño. Selección de subconjuntos de variables en modelos estadísticos. Trabajo Fin de Grado, Universidad de Sevilla. Disponible en https://hdl.handle.net/11441/142918, 2022.

J.D. Kalbfleisch y R.L. Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

W. Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.

D.G. Kleinbaum y M. Klein. *Survival Analysis: A Self-Learning Text*. Springer, Third edition, 2012. doi: https://doi.org/cmnq7s.

E.L. Korn y R. Simon. Measures of explained variation for survival data. *Statistics in medicine*, 9(5):487–503, 1990.

H. Kuhn y A. Tucker. Nonlinear programming. In *Proceedings of 2nd Berkeley symposium*, páginas 481–492. Berkeley: University of California Press, 1951.

S. Lemeshow, S. May y D.W. Hosmer Jr. *Applied survival analysis: regression modeling of time-to-event data*. John Wiley & Sons, 2008. doi: https://doi.org/ftds85.

J. Li y S. Ma. *Survival analysis in medicine and genetics*. CRC Press, 2013.

X. Liu. *Survival analysis: models and applications*. John Wiley & Sons, 2012. doi: https://doi.org/kcd7.

J.S. López. Selección de modelos de supervivencia para pacientes con cáncer gastrointestinal-un enfoque bajo modelos de fragilidad. Trabajo Fin de Máster, Universidad de los Andes. Disponible en http://hdl.handle.net/1992/12348, 2014.

D.G. Luenberger y Y. Ye. *Linear and nonlinear programming*, volume 2. Springer, 1984.

J. Ma. Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction. *IEEE Transactions on Nuclear Science*, 57(1): 181–192, 2010.

D.F. Moore. *Applied survival analysis using R*, volume 473. Springer, 2016.

N.J.D. Nagelkerke. A note on a general definition of the coefficient of determination. *biometrika*, 78(3):691–692, 1991.

J. O'Quigley, R. Xu y J. Stare. Explained randomness in proportional hazards models. *Statistics in medicine*, 24(3):479–489, 2005.

A.L. Palmer. Modelo de regresión de Cox: ejemplo numérico del proceso de estimación de parámetros. *Psicothema*, páginas 387–402, 1993.

A. Pettitt y I.B Daud. Investigating time dependence in Cox's proportional hazards model. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 39(3):313–329, 1990.

M.D. Pérez, J. Basulto y J.A. Camúñez. Un antecedente histórico de regresión lineal: la estimación mediana propuesta por Boscovich. *Gaceta de la Real Sociedad Matematica Española*, 22(2):351–364, 2019.

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

J. Ramírez, E. Regino y S.Y. Guerrero. Comparación de métodos de estimación en regresión de Cox. *Comunicaciones en Estadística*, 10(1):101–112, 2017.

James O Ramsay. Monotone regression splines in action. *Statistical science*, páginas 425–441, 1988.

P. Royston. Explained variation for survival models. *The Stata Journal*, 6(1):83–96, 2006.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020. Disponible en: http://www.rstudio.com/.

G. Ruiz et al. Los orígenes del método de mínimos cuadrados. *Suma*, 2003.

N.M. Russo. Aplicaciones del análisis de supervivencia en la biotecnología. Trabajo Fin de Grado, Universidad de Almería. Disponible en http://hdl.handle.net/10835/13488, 2021.

B. San José, E. Pérez y R. Madero. Métodos estadísticos en estudios de supervivencia. *Anales de Pediatría Continuada*, 7(1):55–59, 2009.

J.P. Sánchez. Modelos semiparamétricos en el análisis de supervivencia aplicado a la mejora genética. *Dep. de Ciencia Animal, UPV*, 2005.

M. Schemper. The explained variation in proportional hazards regression. *Biometrika*, 77 (1):216–218, 1990.

M. Schemper y J. Stare. Explained variation in survival analysis. *Statistics in medicine*, 15(19):1999–2012, 1996.

D. Schoenfeld. Partial residuals for the proportional hazards model. *Biometrika*, 69:51–55, 1982.

M. Spreafico, F. Ieva y M. Fiocco. Modelling time-varying covariates effect on survival via functional data analysis: application to the MRC BO06 trial in osteosarcoma. *Statistical Methods & Applications*, 32(1):271–298, 2023.

J. M. Stanton. Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9(3), 2001. doi: https://doi.org/gd82dx.

M. Tableman y J.S. Kim. *Survival analysis using S: analysis of time-to-event data*. CRC press, 2003.

K. Tay, N. Simon, J. Friedman, T. Hastie, R. Tibshirani y B. Narasimhan. Regularized Cox Regression, 2023.

M. Thackham y J. Ma. On maximum likelihood estimation of the semi-parametric Cox model with time-varying covariates. *Journal of Applied Statistics*, 47(9):1511–1528, 2020.

T. Therneau, C. Crowson y E. Atkinson. Using time dependent covariates and time dependent coefficients in the Cox model. *Survival Vignettes*, 2(3):1–31, 2022.

T.M. Therneau. *A Package for Survival Analysis in R*. R package version 3.5-7, 2023. Disponible en: https://CRAN.R-project.org/package=survival.

T.M. Therneau y P.M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000. doi: https://doi.org/j8kf.

T.M. Therneau, P.M. Grambsch y T.R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

A.A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006. doi: https://doi.org/dx8pkp.

P. Velasco. Modelo de regresión de Cox y sus aplicaciones biosanitarias. Trabajo Fin de Grado, Universidad de Sevilla. Disponible en http://hdl.handle.net/11441/43493, 2016.

A. Webb y J. Ma. Cox models with time-varying covariates and partly-interval censoring–A maximum penalised likelihood approach. *Statistics in medicine*, 42(6):815–833, 2023.

H. Wickham, W. Chang, L. Henry, T.L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani y D. Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.4.3, 2023a. Disponible en: https://CRAN.R-project.org/package=ggplot2.

H. Wickham, R. François, L. Henry, K. Müller y D. Vaughan. *dplyr: A Grammar of Data Manipulation*. R package version 4.1.8., 2023b. Disponible en: https://CRAN.R-project.org/package=glmnet.

W.H. Wong y T.A. Severini. On maximum likelihood estimation in infinite dimensional parameter spaces. *The Annals of Statistics*, páginas 603–632, 1991.

J. Xu, J. Ma, M.H. Connors y H. Brodaty. Proportional hazard model estimation under dependent censoring using copulas and penalized likelihood. *Statistics in medicine*, 37 (14):2238–2251, 2018.

Z. Zhang, J. Reinikainen, K.A. Adeleke, M.E. Pieterse y C.G.M. Groothuis-Oudshoorn. Time-varying covariates and coefficients in Cox regression models. *Annals of translational medicine*, 6(7), 2018.

R. Álvarez. *Sir Francis Galton, padre de la eugenesia*, volume 4. Editorial CSIC-CSIC Press, 1985.