Trabajo final de grado

GRADO DE MATEMÁTICAS

Facultad de Matemáticas y Informática
Universidat de Barcelona

Modelo de Cox de riesgos proporcionales

Autor: Juan Luis Díaz Jiménez

Director: Dr. Olga Julià de Ferran
Realizado en: Departamento
de Matemáticas y Informática

Barcelona, 27 de junio de 2018

Abstract

The objective of this paper is to study the Cox model of proportional risks and apply it to real data to build a model adjusted to them.

First, the basic functions necessary for the analysis of survival, as well as their properties, are introduced.

The censorship and truncation of our data is defined below, phenomena very common in this type of studies given the nature of the data studied.

The next chapter is about the Cox model, how it is defined and its main characteristics. It is explained which parameters compose it and the method of best estimate. The chapter ends with a study of Cox residuals, as well as its importance when it comes to seeing the validity of the adjustment.

Finally, we apply the Cox model to some data on patients with multiple myeloma. From the data of 48 subjects, an adjusted model is elaborated, the residuals are analyzed and it is checked if our adjustment is good. To conclude, some small conclusions are drawn about how each variable influences the risk of death.

Resumen

Este trabajo tiene por objetivo estudiar el modelo de Cox de riesgos proporcionales y aplicarlo a datos reales para construir un modelo ajustado a ellos.

Primeramente se introducen las funciones básicas necesarios para le análisis de la supervivencia, así como sus propiedades.

A continuación se define la censura y el truncamiento de nuestros datos, fenómenos muy comunes en este tipo de estudios dada la naturaleza de los datos estudiados.

El siguiente capítulo trata sobre el modelo de Cox, como se define y sus características principales. Se explica que parámetros lo componen y el método de mejor estimación. El capítulo finaliza con un estudio de los residuos de Cox, así como su importancia a la hora de ver la validez del ajuste.

Por último aplicamos el modelo de Cox a unos datos sobre pacientes con mieloma múltiple. A partir de los datos de 48 inidviduos se elabora un modelo ajustado, se analizan los residuos y se comprueba si ha sido un buen ajuste o no. Para finalizar se extraen unas pequeñas conclusiones sobre como influye cada variable en el riesgo de muerte.

Agradecimientos

Quiero agradecer este trabajo a mi tutora con la que tantas mañanas nos hemos peleado con R y a mis padres y mi hermana que siempre me han apoyado y ayudado con mis estudios.

Índice

1. Introducción

Los orígenes del análisis de supervivencia los encontramos en las tablas de vida, tablas en las que se tabula los tiempos de muerte de una población en función de la edad. El propósito de estas tablas era conocer el patrón de mortalidad de la población.

El gran desarrollo del análisis de supervivencia se experimentará a partir de la segunda mitad del siglo XX cuando este comienza a aplicarse a estudios clínicos. A partir de este momento los estudios médicos y el análisis de supervivencia estarán estrechamente ligados. Sin embargo eso no quiere decir que no tenga muchas más aplicaciones ya que hoy en día puede aplicarse a procesos industriales (análisis de fiabilidad), procesos económicos y sociológicos (análisis de duraciones), estudios demográficos (análisis de la historia de los sucesos), etc.. En definitiva, el estudio de la supervivencia será el estudio de los datos que representan la duración entre dos sucesos ya sean el nacimiento y muerte de un individuo, la duración de un matrimonio o el tiempo que tarde en fallar un componente electrónico.

Es importante resaltar que el análisis de supervivencia nos da como resultado el tiempo hasta el suceso de interés. A menudo las características especiales de los tiempos de vida no se modelan con facilidad utilizando distribuciones paramétricas y serán necesarios modelos no paramétricos. Sin embargo estos modelos también presentan algún inconveniente ya que suelen dar estimadores excesivamnete imprecisos y la interpretación de algunas funciones queda limitada debido a su discontinuidad.

El modelo de Cox, o modelo de riesgos proporcionales es uno de los modelos más utilizados en el análisis de la supervivencia. Este modelo fue propuesto por David Cox en una serie de trabajos, el más importante de ellos es *Regression Models and Life Tables* publicado en 1972 en el *Journal of the Royal Statistical Society (serie B)*. Este modelo se considera semi-paramétrico debido a que combina una función paramétrica con otra función que puede tomar cualquier forma, como veremos en su capítulo correspondiente.

2. Capítulo 1. Conceptos básicos

Denominamos T el tiempo que transcurre hasta el suceso de interés, denominado ε. Es una variable aleatoria no negativa que corresponde a una población homogénea. T puede ser tanto continua o discreta como resultado de redondear o agrupar tiempos en intervalos.

El suceso ε puede ser la muerte del paciente, la aparición o desaparición de una enfermedad, el fallo mecánico de una máquina, etc.

El modelo de T se puede caracterizar por seis funciones, la de supervivencia, distribución, densidad de probabilidad, de riesgo, de riesgo acumulado y vida media residual. Para nuestro estudio nos interesarán sobre todo la función de supervivencia y la función de riesgo aunque también definiremos la función densidad ya que es necesaria para demostrar algunas propiedades de la función de riesgo.

2.1. Función de supervivencia

La función de supervivencia se define como la probabilidad de que un individuo sobreviva más de t unidades de tiempo (por ejemplo días). También se puede interpretar como la probabilidad de que el suceso ε pase después de t (días).

$$S(t) = \text{Prob}(T > t), \tag{2.1}$$

definida para todo t no negativo.

Algunas de sus propiedades son

1. $S(0) = 1$ y $\lim_{t\to\infty} S(t) = 0$.
2. Es monótona decreciente.
3. Si T es continua, $S(t)$ es continua y estricamente decreciente.

Las funciones de supervivencia puede tomar muchas formas pero todas parten de 1 y decrecen monótonicamente y convergen a 0 cuando t tiende a ∞.

2.2. Función de densidad de probabilidad

La función densidad de probabilidad para variables absolutamente continuas se define como

$$f(t) = \lim_{\Delta t\to 0} \frac{1}{\Delta t}\text{Prob}[t \le T < t + \Delta t]. \tag{2.2}$$

Esta función mide la intensidad de la probabilidad y $f(t)\Delta t$ se interpreta como la probabilidad de que el evento ε ocurra entre t y $t + \Delta t$.

Las propiedades básicas de la función de densidad son

1. $f(t)$ es una función no negativa.

2. El área bajo la curva de $f(t)$ vale 1.

3. La función de supervivencia es la integral de la función densidad

$$S(t) = \text{Prob}(T > t) = \int_t^\infty f(u)du$$

4. $f(t)$ también puede interpretarse como el negativo de la velocidad de cambio de la función de supervivencia

$$f(t) = \frac{-dS(t)}{dt}$$

Si las variables son discretas se define la función de masa de probabilidad como

$$p(t_j) = \text{Prob}\{T = t_j\}, \forall j \tag{2.3}$$

Las propiedades básicas de la función de masa de probabilidad son

1. Es una función no negativa.

2. El sumatorio de las masas es igual a 1.

3. $S(t) = \text{Prob}\{T > t\} = \sum_{t_j > t} p(t_j)$.

2.3. Función de riesgo

La función de riesgo cuando t es una variable aleatoria absolutamente continua se define como

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \text{Prob}[t \leq T < t + \Delta t | T \geq t]. \tag{2.4}$$

Es una definición similar a la densidad pero la probabilidad es condicionada a que el evento ε no se ha producido antes de t.

Las propiedades básicas de la función supervivencia cuando T es absolutamente continua son

1. $\lambda(t)$ es una función no negativa.

2. Para s positivo no nulo $\int_0^s \lambda(u)du < \infty$ y $\int_0^\infty \lambda(u)du = \infty$.

3. $\lambda = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\ln(S(t)))$. Por lo tanto podemos rescribir la función supervivencia como $S(t) = e^{-\int \lambda(t)dt}$.

Veamos la demostración de la última propiedad.

Demostración.

$$\lambda(t) = \lim_{\Delta\to 0} \frac{1}{\Delta t}\text{Prob}[t \le T < t + \Delta t | T \ge t] = \lim_{\Delta\to 0} \frac{\text{Prob}[t \le T < t + \Delta t]}{\Delta t \text{Prob}[T \le t]}$$
$$= \frac{f(t)}{S(t)} = \frac{-dS(t)/dt}{S(t)} = -\frac{d}{dt}[\log S(t)]$$

□

Si T es discreta y toma valores $t_1 < t_2 < ...$ la función de riesgo se define como

$$\lambda(t_j) = \text{Prob}[T = t_j | T \ge t_j] = \text{Prob}[T = t_j | T > t_{j-1}] \tag{2.5}$$

Las propiedades básicas cuando T és discreta son

1. $\lambda(t_j) = \frac{p(t_j)}{S(t_{j-1})} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}$
2. La función de supervivencia puede escribirse como producto de 1 menos la función de riesgo.

$$S(t) = \prod_{t_j \le t} \frac{S(t_j)}{S(t_{j-1})} = \prod_{j:t_j \le t} (1 - \lambda(t_j)). \tag{2.6}$$

Demostración.

$$S(t_j) = \text{Prob}(T > t_j) = \text{Prob}(T > t_j, T > t_{j-1})$$
$$= \text{Prob}(T > t_j | T > t_{j-1})\text{Prob}(T > t_{j-1})$$
$$= (1 - \text{Prob}(T \le t_j | T > t_{j-1}))S(t_{j-1}) = (1 - \lambda(t_j))S(t_{j-1})$$

Recursivamente

$$S(t) = \prod_{t_j \le t} \frac{S(t_j)}{S(t_{j-1})} = \prod_{j:t_j \le t} (1 - \lambda(t_j)).$$

□

La función de riesgo describe el comportamiento de la probabilidad de sobrevivir en un pequeño intervalo de tiempo teniendo en cuenta que la persona está viva al inicio de este mismo.

Si la supervivencia decae rápido el riesgo es alto, por el contrario si la curva de la supervivencia es plana el riesgo es cero.

2.3.1. Tipos de función de riesgo

Las funciones de riesgo pueden ser muy distintas. En este caso distinguiremos cinco tipos:

1. Función de riesgo creciente: el riesgo aumenta con el tiempo, hay efecto envejecimiento.

2. Función de riesgo constante: no hay ni envejecimiento ni fortalecimiento.

3. Función de riesgo decreciente: los individuos se fortalecen con el tiempo.

4. Función de riesgo con forma de bañera: corresponde a la vida de los seres biológicos: primero hay la mortalidad infantil (la curva decrece) luego un periodo de juventud (curva constante) y luego el envejecimiento (curva creciente).

5. Función de riesgo en forma de joroba: la encontramos cuando estudiamos eventos como infecciones después de una operación o rechazo de un trasplante, en que el evento tarda en aparecer pero pasado un tiempo cada vez es menos probable.

2.3.2. Función de riesgo acumulada

La función de riesgo acumulada para T absolutamente continua se define como

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{2.7}$$

Si T es discreta la función de riesgo acumulada es

$$\Lambda(t) = \sum_{j:t_j \leq t} \lambda(t_j) \tag{2.8}$$

Está función carece de interpretación, nos sirve como herramienta matemática.

3. Capítulo 2. Censura y truncamiento

Llamamos censura cuando al realizar un análisis de supervivencia la información de la supervivencia de algunos sujetos no es completa, es decir, el tiempo en que ocurre el suceso ε no se observa, ya sea porque el estudio se termina antes de que ε ocurra, el sujeto abandone el estudio, etc.

Hay varios tipos de censura, pero cuando se realizan estudios con el modelo de Cox la censura que presentaran nuestros datos más típica será la censura por la derecha.

3.1. Censura por la derecha

Se sigue a los sujetos hasta el momento en que ocurre ε. Si éste ocurre entonces el tiempo hasta ε es conocido. Si al finalizar el estudio el suceso no ha ocurrido entonces la observación se dice que está censurada a la derecha. Esta puede darse por:

1. El estudio termina en un momento predeterminado. Por ejemplo en ensayos clínicos o en estudios sobre máquinas hechos antes de que todas las máquinas fallen.

2. Perdida de seguimiento de algunos sujetos debido por ejemplo a un cambio de domicilio o hospital.

3. Interrupción del tratamiento. El sujeto ya no podría participar del estudio.

4. El suceso ocurre por una causa distinta a la de interés. Por ejemplo si el suceso de interés es la muerte del sujeto a causa de un cáncer de pulmón pero este muere debido a un accidente de tráfico.

La censura es no informativa si el hecho de conocer el tiempo de censura de un sujeto no aporta más información que sobre la supervivencia del mismo que la que se tendría si hubiese continuado el estudio. Si el tiempo de supervivencia y el tiempo de censura son independientes entonces la censura es no informativa.

Los tipos de censura por la derecha son censura de tipo I fija, progresiva o generalizada, censura tipo II y censura aleatoria. Esta última es la más común en los estudios donde se utiliza el modelo de Cox y es por eso que se explicará con más detalle.

3.1.1. Censura aleatoria

Habrá ocasiones en las que los sujetos abandonarán el estudio debido a alguna de las causas anteriormente enumeradas. Si la censura es independiente de los tiempos de muerte el análisis se simplifica.

Si tenemos una muestra de n sujetos con tiempos potenciales hasta ε $T_1, ..., T_n$ censurados por la derecha observaremos $(Y_1, \delta_1), ..., (Y_n, \delta_n)$ con Y_i =min$\{T_i, C_i\}$, donde $C_1, ..., C_n$ son los tiempos de censura de cada sujeto y $\delta_i = 1$ si $T_i \leq C_i$ (el sujeto i no está censurado) o $\delta_i = 0$ si $T_i < C_i$ (el sujeto i está censurado).

3.2. Truncamiento

Decimos que los individuos presentan truncamiento si estos han sidos filtrados de tal forma que no podemos saber de su existencia. La diferencia con la censura, que nos permite saber la información de un individuo hasta el tiempo censurado, es que un individuo truncado no sabemos que existe. El truncamiento puede ser tanto por la derecha como por la izquierda.

Un ejemplo de truncamiento por la derecha es cuando solo se toman individuos para el estudio que hayan experimentado el suceso ϵ. Y un ejemplo de truncamiento por la izquierda es cuando solo se toman individuos hasta una cierta edad.

4. Capítulo 3. Modelo de Cox de riesgos proporcionales

Denominamos T una variable aleatoria que indica el tiempo hasta un suceso ε. Sean $Z_1, ..., Z_p$ covariantes fijas (variables explicativas tomadas al inicio del estudio). El conjunto de covariantes asociadas a un individuo o factores pronósticos se denotan por $Z = (Z_1, Z_2, ..., Z_p)$ y se denomina el perfil del sujeto.

Sea $\lambda_0(t)$ la función de riesgo basal. Dicha función es la que le correspondería a un individuo cuyo perfil fuese $Z = 0$ y puede tomar cualquier forma. Es por esto que el modelo de Cox es semiparamétrico.

El modelo de Cox escribe la función de riesgo de un sujeto con perfil Z en un instante t en función de la función de riesgo basal multiplicada por una función no negativa que se interpreta como el riesgo relativo en el instante t.

$$\lambda(t|Z) = \phi(Z)\lambda_0(t) \tag{4.1}$$

ϕ podemos parametrizarla de tres formas:

1. log-lineal: $\phi(Z;\beta) = e^{\beta Z}$
2. lineal: $\phi(Z;\beta) = 1 + \beta Z$
3. logística: $\phi(Z;\beta) = ln(1 + e^{\beta Z})$

La más común es la parametrización log-lineal debido a las buenas propiedades que presenta para trabajar. El modelo básico de Cox establece la siguiente regresión multivariada:

$$\lambda(t|Z) = e^{\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p}\lambda_0(t) \tag{4.2}$$

Donde la función basal es común a todos los sujetos y la función de riesgo relativo esta determinada por los coeficientes de regresión $\beta_1, ..., \beta_p$ y las covariantes fijas de cada individuo.

Es imporatne notar que este modelo supone que la razón entre las funciones de riesgo es constante a lo largo del tiempo ya que $e^{\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p}$ solo depende de las covariantes. Este factor corresponde al riesgo entre un sujeto con perfil Z y otro con $Z = 0$ y nos indica cuantas veces es mayor el riesgo de que suceda ε a un sujeto con perfil Z que con $Z = 0$.

4.1. Modelo de Cox y datos censurados por la derecha

Es posible que, debido a las características de nuestro estudio, nuestros datos presenten algún tipo de censura. La más común y en la que nos centraremos será la censura por la derecha aleatoria. Como hemos visto en el apartado (3.1.1) nuestros datos vendrán dados por (Y_i, δ_i). A este par añadiremos el vector de covariantes fijas $Z_i = (Z_{1i,2i,...,pi})$ obteniendo la terna (Y_i, δ_i, Z_i) para el individuo i.

Hemos de suponer que T_i y Z_i cumplen con la ecuación (4.2) y las variables aleatorias T_i y C_i son condicionalmemte independientes para un Z_i dado.

4.2. Variables continuas y factores

En nuestro estudio las covariantes fijas las dividiremos en dos tipos: Variables y factores.

1. Variables: Toman valores numéricos normalmente en una escala continua y se añaden al modelo asignando a cada variable un coeficiente β.

2. Factores: Toman un número limitado de valores a los que llamamos niveles de factor. Para incorporar un factor F con k niveles añadiremos $k-1$ coeficientes α_j, que se definen como los efectos principales del factor F, en el modelo y supondremos que el primer nivel es el nivel de referencia. Para poder trabajar con estos factores en el programa informático que vayamos a utilizar (en nuestro caso será R) crearemos unas variables *dummy*. Si tenemos k niveles necesitaremos $Z_2, ..., Z_k$ variables.

4.3. Estimación de los coeficientes β

Como vemos por la ecuación (4.2) el modelo de Cox esta determinado por la función basal y los coeficientes de las variables fijas. Estos coeficientes no es posible calcularlos de forma analítica, por lo tanto lo que haremos será calcular una estimación a partir de la maximización de la función de verosimilitud parcial $L(\beta)$.

Una vez tengamos los estimadores calculados podremos estimar la función basal aunque para la mayoría de estudios no será necesario ya que lo que más nos interesará será el estudio comparado de diferentes individuos. Como la función basal es idéntica para todos los individuos no la tendremos en cuenta.

4.3.1. Función de verosimilitud parcial

Antes de poder definir $L(\beta)$ necesitamos introducir una serie de nuevos elementos. Sea r el número de tiempos que llegan hasta el suceso de interés (a partir de este momento lo denominaremos por muerte) y $n-r$ el número de tiempos censurados (suponemos que $r \neq n$). Ordenamos los tiempos r y los denominamos $t_{(1)}, ..., t_{(r)}$. Denominamos $R_j = R(t_{(j)}) = \{i : Y_i \geq t_{(j)}\}$ al conjunto de individuos que están en riesgo de morir (siguen vivos y no están censurados) en el momento $t_{(j)}$.

Denominamos $\Gamma = \{(Y_i, \delta_i, Z_i), i = 1, ..., n\}$ al conjunto que contiene toda la información de la muestra y Γ_j al conjunto con toda la información hasta el momento $t_{(j)}$. Sea $Z_{(j)}$ el vector de covariantes aleatorias del individuo que ha muerto en el tiempo $t_{(j)}$.

Denominamos $e_1, e_2, ..., e_r$ las etiquetas que identifican a los individuos con tiempos de muerte $t_{(1)}, ..., t_{(r)}$. Denotamos $Z_{1j}, z_{2j}, ..., z_{nj}$ las covariantes de losn_j individuos del conjunto de riesgo $R(t_{(t)})$.

Ejemplo 4.1. Tenemos $n = 5$ individuos.

Individuo	Y	δ	Z
1	21	0	Z_1
2	6	1	Z_2
3	15	1	Z_3
4	7	0	Z_4
5	10	1	Z_5

Cuadro 1: Ejemplo de cinco individuos.

Podemos ver que hay dos individuos censurados, entonces $r = 5 - 2 = 3$. Los $\{t_{(1)}, t_{(2)}, t_{(3)}\} = \{6, 10, 15\} = \{Y_2, Y_5, Y_3\}$. Los conjuntos de riesgo son $R_1 = \{1, 2, 3, 4, 5\}, R_2 = \{1, 3, 5\}, R_3 = \{1, 3\}$.

Teorema 4.2. *Supongamos que no tenemos censura, es decir $r = n$, entonces*

$$Prob\{e_1 = i_1, e_2 = i_2, ..., e_n = i_n\} = \prod_{j=1}^{n} \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_{lj}}} \tag{4.3}$$

Demostración. En el modelo de Cox, para poder estimar los β no es necesario conocer el valor de los tiempos de fallo si no tan solo conocer su orden, ya que la función $\lambda_0(t)$ es una función desconocida.

$$\text{Prob}\{e_1 = i_1, e_2 = i_2, ..., e_n = i_n\} = \int_0^\infty \int_{t_{(1)}}^\infty \cdots \int_{t_{(n-1)}}^\infty \prod_1^n f(t_{(i)}; z_{(i)}) dt_{(n)} dt_{(n-1)} \dots dt_{(1)} \tag{4.4}$$

Utilizando que $f(t) = \lambda(t)S(t)$ y $S(t) = e^{-\int \lambda(t)dt} = e^{-\Lambda(t)}$ podemos reescribir las integrales de la ecuación 4.4

$$\int_0^\infty \int_{t_{(1)}}^\infty \cdots \int_{t_{(n-1)}}^\infty \prod_1^n \lambda(t_{(i)}; z_{(i)}) e^{-\Lambda(t_{(i)}; z_{(i)})}$$

Como estamos trabajando con el modelo de Cox $\lambda(t) = \lambda_0(t)e^{\beta' z}$ y $\Lambda(t) = \Lambda_0(t)e^{\beta' z}$. Por lo tanto

$$\int_0^\infty \int_{t_{(1)}}^\infty \cdots \int_{t_{(n-1)}}^\infty \prod_1^n \lambda_0(t_{(i)}) e^{z_{(i)}} e^{-\Lambda_0(t_{(i)}) e^{z_{(i)}}}$$

$$= \prod_{j=1}^{n} \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_{lj}}}.$$

□

Si tenemos censura ($r \neq n$), en este caso

$$\text{Prob}\{e_1 = i_1, e_2 = i_2, ..., e_r = i_r\} \propto \prod_{j=1}^{r} \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_{lj}}} \tag{4.5}$$

La función de verosimilitud parcial se define como el producto de la probabilidad condicional de que el individuo con $z_{(j)}$ se muera en $t_{(j)}$ sabiendo qe ha ocurrido una muerte entre los n_j individuos en riesgo en el momento $t_{(j)}$.

$$L(\beta_1, ..., \beta_p) = \prod_{j=1}^{r} \text{Prob}\{e_j = i | \Gamma_j\} = \prod_{j=1}^{r} \text{Prob}\{Z_{(j)} = z_{(j)} | \Gamma_j\} \tag{4.6}$$

Por lo tanto la función de verosimilitud parcial es

$$L(\beta_1, ..., \beta_p) = \prod_{j=1}^{r} \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_{lj}}} \tag{4.7}$$

Ejemplo 4.3. Siguiendo con el ejemplo 4.1 veamos ahora cual es su función de verosimilitud parcial.

$$L(\beta_1, ..., \beta_p) = \prod_{j=1}^{3} \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_{lj}}}$$

$$= \frac{e^{\beta' z_2}}{e^{\beta' z_1} + e^{\beta' z_2} + e^{\beta' z_3} + e^{\beta' z_4} + e^{\beta' z_5}} \cdot \frac{e^{\beta' z_5}}{e^{\beta' z_1} + e^{\beta' z_3} + e^{\beta' z_5}} \cdot \frac{e^{\beta' z_3}}{e^{\beta' z_1} + e^{\beta' z_3}} \tag{4.8}$$

4.3.2. Maximización de la función de verosimilitud

La función de verosimilitud la podemos escribir de tal manera que en el numerador dependa de los individuos que han experimentado el suceso y el denominador dependa de aquellos que no lo han experimentado todavía.

$$L(\beta_1, ..., \beta_p) = \prod_{j=1}^{r} \frac{e^{\sum_{k=1}^{p} \beta_k z_{(j)k}}}{\sum_{l \in R(t_{(j)})} e^{\sum_{k=1}^{p} \beta_k z_{jk}}} \tag{4.9}$$

Denominamos $\hat{\beta} = (\hat{\beta}_1, ...; \hat{\beta}_p)$ al estimador de los coeficientes β. Este se obtiene maximizando la función de verosimilitud parcial mediante métodos numéricos.

Teorema 4.4. *Para un número suficientemente grande se cumple:*

1. *$\hat{\beta}$ estima consistentemente β, es decir, $\hat{\beta}$ converge a β cuando el numero de datos tiende a infinito.*

2. *La distribución es aproximadamente normal.*

3. *El estimador $\hat{\beta}$ no es eficiente, es decir, no alcanza la cota de Cramer-Rao.*

4. Los estimadores no son independientes entre ellos.

Para determinar si algún coeficiente β_j es 0 utilizamos el cálculo del *p-value* y el intervalo de confianza

$$[\hat{\beta}_j - z_{\frac{\alpha}{2}}\sqrt{\hat{\text{Var}}\hat{\beta}_j}, \hat{\beta}_j + z_{\frac{\alpha}{2}}\sqrt{\hat{\text{Var}}\hat{\beta}_j}]. \tag{4.10}$$

Que el intervalo de confianza incluya el 0 o su *p-value* sea mayor que $0,05$ puede indicarnos que el valor de este coeficiente sea 0 y por lo tanto su covariante asociada no influya en el modelo. Aun así no podemos excluirla de inmediato, ya que como hemos dicho anteriormente los β_j no son independientes entre ellos. Al quitar uno, el valor del resto puede cambiar. Para saber cual es el mejor subconjunto de covariantes que explica el modelo utilizaremos la comparación de modelos anidados.

4.3.3. Comparación de modelos anidados

Tenemos $p+q$ covariantes $Z_1, ..., Z_p, Z_{p+1}, ..., Z_{p+q}$ y dos modelos. El primero solo tiene en cuenta las p primeras covariantes y el segundo las $p+q$.

El segundo modelo ajusta mejor nuestros datos que el primero ya que contiene toda la información del primero más términos adicionales. Lo que nos interesa saber es si las q variables adicionales mejoran significativamente el primer modelo. Si no fuese así concluiríamos que el primer modelo es más adecuado.

Como ya se ha dicho antes cada coeficiente depende del resto, por lo tanto no podemos estudiarlos de forma independiente, por ejemplo el coeficiente de Z_i en el segundo modelo depende de las $p+q-1$ variables con las que ha sido ajustado. Es por eso que diremos que el efecto de Z_i está ajustado por $Z_1, ..., Z_{i-1}, Z_{i+1}, ..., Z_{p+q}$.

Los criterios que utilizaremos para saber que modelo ajusta mejor nuestros datos serán dos, el estadístico deviance y el criterio de Akaike ambos basados en el valor de la verosimilitud.

La deviance se define como

$$D = -2\text{log}(\hat{L}(\hat{\beta})). \tag{4.11}$$

La deviance siempre toma valores positivos. Cuanto mayor sea este valor mejor ajustado será el modelo. Sean D_1 y D_2 la deviance de un modelo 1 y 2, la diferencia entre ambas es

$$\begin{aligned} D_1 - D_2 &= -2\text{log}(\hat{L_1}(\hat{\beta})) - (-2\text{log}(\hat{L_2}(\hat{\beta}))) \\ &= -2\text{log}(\hat{L_1}(\hat{\beta})) + 2\text{log}(\hat{L_2}(\hat{\beta})) = -2\text{log}(\frac{\hat{L_1}(\hat{\beta})}{\hat{L_2}(\hat{\beta})}). \end{aligned} \tag{4.12}$$

El criterio de Akaike utiliza el estadístico AIC

$$\text{AIC} = -2\text{log}((\hat{L}(\hat{\beta})) + kp \tag{4.13}$$

donde p es el numero de regresores y k una constante que normalmente se toma igual a 2. El valor del AIC decrece conforme se añaden variables al modelo hasta un punto en el que comienza a crecer. Es esto lo que nos indica que ya tenemos suficientes variables. El criterio de Akaike será el que se utilizará para la construcción de nuestro modelo en el capítulo siguiente, ya que es el más equilibrado entre un modelo que ajuste bien los datos y uno que contenga pocas variables.

4.3.4. Construcción del modelo

Para construir nuestro modelo seguiremos estos pasos:

1. Calcular la estimación del coeficiente, su *p-value* y estadístico AIC para cada covariante de forma individual.
2. Tomaremos la covariante con el *p-value* y AIC menor.
3. Volvemos a calcular los coeficientes, esta vez en lugar de forma individual añadiremos a la covariante seleccionada en el paso anterior de forma que ahora el modelo contendrá dos covariantes.
4. Tomaremos la covariante con el *p-value* menor y que haya hecho disminuir el AIC con respecto al paso 2.
5. Repetiremos este procedimiento añadiendo covariantes al modelo hasta que el valor del AIC solo aumente.

Una vez hayamos seguido estos pasos habremos descartado las covariantes menos relevantes para nuestro modelo estadísticamente hablando. Aun así es importante tener en cuenta los aspectos no estadísticos de nuestro estudio. En ocasiones también puede ser necesario aplicar algún tipo de transformación de las variables o considerar tendencias no lineales, aunque esto lo veremos mejor con el estudio de los residuos.

4.4. Estimación de la función de riesgo y supervivencia

Una vez tenemos los estimadores de los coeficientes β podemos estimar cual es la probabilidad que un individuo con un vector de covariantes Z tiene de sobrevivir un tiempo t. La estimación de la función de riesgo y supervivencia para el modelo de Cox de riesgos proporcionales fue desarrollada por Kalbfleisch y Prentice en 1973.

4.4.1. Estimadores de Kalbfleisch y Prentice

Para cada tiempo de fallo $t_{(j)}$ definimos los parámetros $\xi_j = 1 - \lambda_0(t_{(j)})$ para $j = 1, ..., r$. Si suponemos que la funció de reisgo es constante entre dos tiempos de muerte consecutivos entonces $\xi_j = \text{Prob}(T \geq t_{(j+1)} | T \geq t_{(j)})$ (probabilidad que un individuo sobreviva al intervalo $[t_{(j)}, t_{(j+1)})$.

Si suponemos que no hay empates el estimador de máxima verosimilitud para ξ es

$$\hat{\xi}_j = \left(1 - \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_{(l)}}}\right) \tag{4.14}$$

El estimador de máxima verosimilitud de $S_0(t)$ es una función escalonada que se expresa como

$$\hat{S}_{0KP}(t) = \prod_{j=1}^{k} \hat{\xi}_j \text{ para } t_{(k)} \leq t < t_{(k+1)}, k = 1, ..., r-1. \tag{4.15}$$

Para todo $t > t_{(r)}$ $\hat{S}_{0KP}(t) = 0$ si no tenemos datos censurados mayores de $t_{(r)}$. Si existen datos censurados mayores que $t_{(r)}$ la función no está definida para ellos.

El estimador de la función de riesgo subyacente en el momento $t_{(j)}$ es

$$\hat{\lambda}_{0KP}(t_{(j)}) = 1 - \hat{\xi}_j. \tag{4.16}$$

La función de riesgo subyacente acumulada se puede estimar mediante

$$\hat{\Lambda}_{0KP1}(t) = -\log \hat{S}_{0KP}(t) = -\sum_{j=1}^{k} \log \hat{\xi}_j, \tag{4.17}$$

o como

$$\hat{\Lambda}_{0KP2}(t) = \sum_{j=1}^{k} (1 - \hat{\xi}_j). \tag{4.18}$$

Una vez tenemos calculado estos estimadores podemos escribir los estimadores de las funciones de riesgo, riesgo acumulada y supervivencia de un individuo con un vector de covariantes Z_i como:

$$\hat{\lambda}(t|Z_i) = \hat{\lambda}_{0KP}(t_{(j)}) e^{\hat{\beta}' Z_i} \tag{4.19}$$

$$\hat{\Lambda}(t|Z_i) = \hat{\Lambda}_{0KP1}(t) e^{\hat{\beta}' Z_i} \text{ ó } \hat{\Lambda}(t|Z_i) = \hat{\Lambda}_{0KP2}(t) e^{\hat{\beta}' Z_i} \tag{4.20}$$

$$\hat{S}(t|Z_i) = [\hat{S}_{0KP}(t)]^{e^{\hat{\beta}' Z_i}}. \tag{4.21}$$

4.5. Análisis de los residuos en el modelo de Cox

Para saber si nuestro modelo se ajusta bien a los datos del estudio será necesario realizar un análisis de los residuos. La definición de éstos no es única y distintas definiciones darán lugar a estudios de diferentes aspectos a la hora de comprobar la validez de nuestro modelo.

4.5.1. Residuos de Cox-Snell

El residuo de Cox-Snell para el i-ésimo individuo se define como

$$r_{C_i} = e^{\hat{\beta}' Z_i \hat{\Lambda}_0(y_i)}, i = 1, ..., n \tag{4.22}$$

donde $\hat{\Lambda}_0(y_i)$ es el estimador de Nelson-Alen de la función de riesgo acumulada. Estos residuos no se distribuyen de forma simétrica alrededor del 0 y siempre son positivos.

Estos residuos sirven para comprobar el ajuste global del modelo. Si hacemos el gráfico de la función de riesgo acumulada para los r_{C_i}, $\hat{\Lambda}_r(r_{C_i})$, contra cada uno de los r_{C_i} deberíamos obtener una recta que pase por el origen con una pendiente de 45º si el modelo está bien ajustado.

Sin embargo tienen una serie de deficiencias que no los hacen recomendables a la hora de estudiarlos. Si el gráfico no se ajusta a una recta no podemos saber cuan alejado del modelo está nuestro ajuste. Además, la lejanía de la linealidad que pueda presentar el gráfico puede ser debida a la imprecisión en la estimación de los parámetros. Pero sobre todo la mayor desventaja es que estos residuos siempre se distribuyen aproximadamente como una exponencial independientemente que el modelo sea válido o no.

4.5.2. Residuos basados en martingalas

Los residuos basados en martingalas para el i-ésimo individuo se define como

$$r_{M_i} = \delta_i - r_{C_i} = \begin{cases} 1 - r_{C_i} = 1 - e^{\hat{\beta}' Z_i \hat{\Lambda}_0(y_i)} & \text{si la observación no está censurada} \\ -r_{C_i} = -e^{\hat{\beta}' Z_i \hat{\Lambda}_0(y_i)} & \text{si la observación está censurada} \end{cases} \tag{4.23}$$

Los residuos de martingalas son una modificación de los residuos de Cox-Snell. Ahora los residuos toman valores entre $-\infty$ y 1 para las observaciones que no están censuradas y $-\infty$ y 0 para las que si lo están. Al igual que los anteriores residuos estos tampoco se distribuyen simétricamente alrededor de 0.

Los residuos de martingalas se pueden interpretar como la diferencia entre el número de eventos observados y el esperado por el modelo de Cox. Sus propiedades son:

1. El valor esperado de cada residuo evaluado en β es 0.
2. La suma de todos los residuos calculados para $\hat{\beta}$ es nula.
3. Los residuos calculados en β están no correlacionados.
4. Para $\hat{\beta}$ los residuos están correlacionados negativamente.

Los residuos de martingalas los utilizaremos para dos cuestiones principalmente:

La primera, es que estos residuos estiman el número de eventos excedidos en los datos pero no predichos por el modelo, es decir, ponen de manifiesto aquellos sujetos que no se ajustan bien al modelo, ya sea porque viven más de la cuenta o porqué mueren muy pronto.

Como ya se señaló al final del apartado 4.3.4 habrá ocasiones en las que una covariante pueda necesitar una transformación. Con los residuos de martingalas podemos encontrar la mejor. Supongamos que tenemos p covariantes y queremos conocer la función que en aplicarla a Z_i mejora como explica su efecto en la supervivencia. Lo que habrá que hacer es ajustar el modelo de Cox para las $Z_1, ..., Z_{i-1}, Z_{i+1}, ..., Z_p$ y despues calcularemos los residuos basados en martingalas para este modelo. Una vez hecho esto haremos el gráfico de los residuos contra los valores de Z_i. El suavizado de la nube de puntos nos dará una idea de la forma de la función que deberemos tomar. Si el suavizado resulta lineal no será necesaria ninguna transformación.

4.5.3. Residuos basados en el *score*

Los residuos basados en el *score* se definen para el individuo i-ésimo y el covariante k-ésimo como

$$\begin{aligned} \bar{Z}_k(t) &= \frac{\sum_{i=1}^n J_i(t) Z_{ik} e^{\beta' Z_i(t)}}{\sum_{i=1}^n J_i(t) e^{\beta' Z_i(t)}} \\ \hat{M}_i(t) &= N_i(t) - \int_0^t J_i(s) e^{\beta' Z_i(s)} d\hat{\Lambda}_0(s) \\ r_{S_{ik}}(t) &= \int_0^t \{Z_{ik}(s) - \bar{Z}_k(s)\} d\hat{M}_i(s) \end{aligned} \tag{4.24}$$

donde $J_i(t) = 1$ si el individuo i-ésimo está en riesgo antes de t y $N - i$ es el proceso que indica si el individuo i-ésimo ha muerto.

El conjunto de residuos basados en los *score* forma una matriz $n \times p$. Cada residuo establece la diferencia entre el valor observado y el esperado de la covariante Z_{ji} del individuo i-ésimo. Lo que obtenemos es el cálculo aproximado para comparar el estimador $\hat{\beta}$ que conseguimos al utilizar todos los datos con el estimador que obtenemos al usar todos los datos menos el j-ésimo individuo.

Los residuos basados en los *score* son útiles para comprobar la hipótesis de riesgos proporcionales. Además si realizamos el gráfico de los $r_{S_{ji}}$ contra las covariantes Z_{ji} nos permiten ver la influencia del i-ésimo individuo en la estimación de los coeficientes. Por lo tanto serán muy útiles para comprobar la influencia de cada individuo en el ajuste global.

4.5.4. Residuos basados en la *deviance*

Los residuos basados en la *deviance* se definen como

$$r_{D_i} = \text{Signo}(r_{M_i}) \sqrt{-2\{r_{M_i} + \delta_i \log(\delta_i - r_{M_i})\}}. \tag{4.25}$$

Podemos observar que estos residuos son una modificación de los residuos basados en martingalas y son nulos cuando estos últimos también valen 0. Cuando r_{M_i} es cercano a 1 el logaritmo hace que el valor aumente, en cambio si r_{M_i} tiene un valor grade negativo el logaritmo lo encoge.

Cuando el modelo es correcto los residuos basados en la *deviance* les corresponde una distribución normal y se distribuyen simétricamente alrededor del 0.

Estos residuos se utilizan, una vez se tiene listo el modelo definitivo, para decidir la bondad del modelo sobre cada individuo. De esta manera podremos ver que observaciones no han sido bien predichas por el modelo ajustado y serán una gran herramienta para comprobar la validez de predicción. Son especialmente útiles para detectar los individuos que viven más o menos de lo esperado por el modelo.

4.5.5. Residuos de Schoenfeld

Los residuos de Schoenfeld están definidos para el k-ésimo evento y el i-ésimo individuo (si no hay empates) como

$$r_{SC_{ik}}(t) = \delta_i J_i(t)\{Z_{ik} - \bar{Z}_k(T_i)\}. \tag{4.26}$$

Podemos observar que los residuos de Schoenfeld son una aproximación de los residuos basados en el *score*. Estos residuos determinan la diferencia entre el valor observado de la covariante en el k-ésimo tiempo de fallo y el valor de la covariante en ese momento.

Al igual que los residuos basados en el *score*, los residuos de Schoenfeld son útiles para determinar la influencia de cada individuo tiene a la hora de estimar los coeficientes.

También son muy útiles para comprobar la hipótesis de proporcionalidad que es la hipótesis de que los riesgos de dos individuos con covariantes diferentes son proporcionales. Si las covariantes cumplen el modelo de Cox, los residuos de Schoenfeld se distribuyen aleatoriamente.

5. Capítulo 4. Estudio de la supervivencia de pacientes con mieloma múltiple mediante el modelo de Cox

Después de ver el desarrollo teórico del modelo de Cox procederemos a aplicarlo sobre unos datos concretos. Los datos corresponden al estudio efectuado en el *Medical Centre of the University of West Virginia (USA)* sobre la supervivencia de pacientes con mieloma múltiple, un tipo de cáncer de la médula ósea que provoca una proliferación anormal de plasmocitos (glóbulos blancos). Esta enfermedad produce anemia, leuconemia, dolores óseos... y por último la muerte. El objetivo és estudiar la asociación entre diferentes variables explicativas y la supervivencia.

5.1. Datos del estudio

Tenemos 48 pacientes de entre 50 y 80 años. Las variables que se recogen son las siguientes:

1. Time: mínimo entre el tiempo hasta la muerte del paciente y el tiempo hasta el final del estudio.
2. Status: Indicador de la censura. Su valor es 1 si el paciente muere por el mieloma, 0 en cualquier otro caso.
3. Age: Edad del paciente.
4. Sex: Su valor es 1 para hombres, 2 para mujeres.
5. BUN: Niveles de urea en sangre.
6. CA: Calcio.
7. HB: Hemoglobina.
8. PC: Porcentaje de glóbulos blancos en la médula ósea.
9. BJP: Presencia de la proteína Bence-Jones. 0 si la proteína está ausente, 1 si está presente.

Veamos una muestra de como se presentan los 10 primeros individuos.

El programa que utilizaremos para analizar estos datos será R. A lo largo de este capítulo se irán indicando los comandos necesarios que se han de introducir a modo de guía para construir un modelo de Cox.

Lo primero que observamos en el cuadro 2 es que los valores de sex y BJP no son numéricos, eso es porqué será R el encargado de asignarle los valores.

	time	status	age	sex	BUN	CA	HB	PC	BJP
1	13	1	66	Male	25	10	14.6	18	Yes
2	52	0	66	Male	13	11	12.0	100	No
3	5	1	53	Female	15	13	11.4	33	Yes
4	40	1	69	Male	10	10	10.2	30	Yes
5	10	1	65	Male	20	10	13.2	66	No
6	7	0	57	Female	12	8	9.9	45	No
7	66	1	52	Male	21	10	12.8	11	Yes
8	10	0	60	Male	47	9	14.0	70	Yes
9	10	1	70	Male	37	12	7.5	47	No
10	14	1	70	Male	40	11	10.6	27	No

Cuadro 2: Cuadro con los 10 primeros pacientes del estudio.

5.2. Estimación de los coeficientes del modelo

Para estimar los coeficientes utilizaremos la función *coxph*. Veamos primero el resultado de introducir todas las covariantes en la función:

```
dades $<-$ coxph(formula = Surv(time, status) $\sim$ age + sex + BUN + CA + HB +
  PC + BJP, data = myeloma).
```

Lo que nos devuelve `dades` lo recogemos en el cuadro 3.

	coef	exp(coef)	se(coef)	z	p
age	-0.02222	0.97802	0.02839	-0.78	0.43369
sexMale	0.22520	1.25257	0.39619	0.57	0.56976
BUN	0.02095	1.02117	0.00613	3.42	0.00064
CA	0.01707	1.01721	0.13297	0.13	0.89788
HB	-0.14028	0.86911	0.06738	-2.08	0.03735
PC	-0.00193	0.99807	0.00661	-0.29	0.77028
BJPNo	0.62660	1.87123	0.43005	1.46	0.14511

Cuadro 3: Estimación de los coeficientes del modelo, así como el estadístico z y *p-value*.

Aquí ya podemos observar que la covariante BUN es la que tiene el menor *p-value* (muy por debajo de 0.05) y la z más alta. Esto es un indicativo de que BUN será relevante para construir nuestro modelo. También podemos ver que tenemos varias covariantes cuyo *p-value* es superior a 0.05 indicando que pueden ser candidatas para ser eliminadas del modelo. Sin embargo, como dijimos en el apartado 4.3.4, los estimadores de los coeficientes no son independientes unos de otros y al quitar uno el valor del resto se ve modificado. Por lo tanto en lugar de ir sustrayendo aquellas covarientes cuyo *p-value* sea muy elevado lo que haremos será aplicar el criterio de Akaike, calcularemos los coeficientes de forma individual así como su AIC.

El cuadro 4 esta construido utilizando la formula `coxph` pero solo pasandole una sola covariante, por ejemplo

	coef	exp(coef)	se(coef)	z	p	AIC
age	0.00983	1.00988	0.02773	0.35	0.72	216.6415
sexMale	-0.0671	0.9351	0.3546	-0.19	0.85	216.7318
BUN	0.01977	1.01997	0.00568	3.48	5e-04	207.6568
CA	-0.0835	0.9199	0.1323	-0.63	0.53	216.3491
HB	-0.1529	0.8582	0.0589	-2.6	0.0094	210.1953
PC	0.00164	1.00164	0.00567	0.29	0.77	216.6846
BJPNo	0.544	1.723	0.388	1.4	0.16	214.6803

Cuadro 4: Estimación de los coeficientes del modelo de forma individual.

```
dadesage $<-$coxph(formula = Surv(time, status) $\sim$ age, data = myeloma).
```

Para calcular el AIC de cada modelo con una sola covariante utilizamos `{AIC(dadesage)`.

Podemos ver que si los estimadores se calculan por separado su valor es distinto al del cuadro 3. Siguiendo el criterio de Akaike tomamos el modelo con el menor AIC, que es el modelo con la covariante BUN.

El siguiente paso es construir un cuadro similar al anterior, pero esta vez calculando los estimadores y los AIC incluyendo la covariante BUN a la formula `coxph`, por ejemplo

```
dadesBUNage $<-$coxph(formula = Surv(time, status) $\sim$ age + BUN,
 data = myeloma).
```

	AIC
age + BUN	209.6481
sexMale +BUN	209.6426
CA + BUN	209.6342
HB + BUN	204.2418
PC + BUN	209.3734
BJPNo + BUN	205.4532

Cuadro 5: Cálculo del AIC para cada modelo.

Como podemos observar que el valor del AIC a disminuido para todos los modelos, pero el modelo con el menor valor es el que incluye al BUN la covariante HB. A continuación volvemos a calcular los AIC esta vez incluyendo BUN y HC.

	AIC
age + BUN + HC	205.7265
sexMale +BUN + HC	205.9082
CA + BUN + HC	206.2404
PC + BUN + HC	206.1139
BJPNo + BUN + HC	203.8956

Cuadro 6: Cálculo del AIC para cada modelo.

De nuevo todos los AIC disminuyen y el modelo con el menor valor es el que incluye la covariante BJP. Volvemos a calcular los AIC añadiendo esta última.

	AIC
age + BUN + HC + BJP	205.4187
sexMale +BUN + HC + BJP	205.6255
CA + BUN + HC + BJP	205.8583
PC + BUN + HC + BJP	205.8937

Cuadro 7: Cálculo del AIC para cada modelo.

Podemos ver en el cuadro 7 que aunque los AIC de cada modelo disminuyen al añadir la covariante BJP ninguno es inferior al AIC del modelo con las covariantes BUN, HC, BJP. Recordemos lo que nos dice el criterio de Akaike, hemos de ir tomando covariantes mientras disminuyan el AIC hasta un punto en que este aumente. Estas serán las covariantes relevantes. En nuestro modelo son BUN, HC y BJP (el nivel de urea en sangre, la hemoglobina y la presendia de la proteína Bence-Jones). Por lo tanto la influencia de la edad, el sexo, el calcio y la presencia de glóbulos blancos en la función de riesgo no será relevante.

Una vez determinadas las covariantes relevantes procedemos a calcular los estimadores de los coeficientes con la formula

```
dadesBUNHCBJP $<-$ coxph(formula = Surv(time, status) $\sim$ BUN + HB + BJP,
data = myeloma)
```

	coef	exp(coef)	se(coef)	z	p
BUN	0.02043	1.02064	0.00594	3.44	0.00059
HB	-0.11478	0.89156	0.06054	-1.90	0.05795
BJPNo	0.61267	1.84535	0.41038	1.49	0.13546

Cuadro 8: Estimación de los coeficientes del modelo para las covariantes BUN, HB, BJP.

La función de riesgo del modelo será de la forma

$$\frac{\lambda(t|Z)}{\lambda_0(t)} = e^{0,02043Z_{BUN}-0,11478Z_{HB}+0,61267Z_{BJP}} = e^{0,02043Z_{BUN}}e^{-0,11478Z_{HB}}e^{0,61267Z_{BJP}} \tag{5.1}$$

Sin embargo con esto no acabamos nuestra construcción del modelo. Ahora es el momento de estudiar los residuos para ver la bondad de ajuste de nuestro modelo.

5.3. Estudio de los residuos

Primeramente comprobaremos si alguna de las covariantes utilizadas necesita transformarse. Para ello utilizaremos los residuos basados en martingalas utilizando en R la formula

`resid(dadesBUNHBBJP, type="martingale").`

Una vez tengamos los residuos calculados construimos el gráfico de las covariantes versus los residuos y añadimos un suavizado a la nube de puntos. Dado que tenemos 48 individuos y solo 36 no están censurados la nube de puntos se verá bastante dispersa.

Los gráficos solo serán necesarios para BUN y HB, ya que la covariante BJP es del tipo factor y por lo tanto no admite transformaciones.

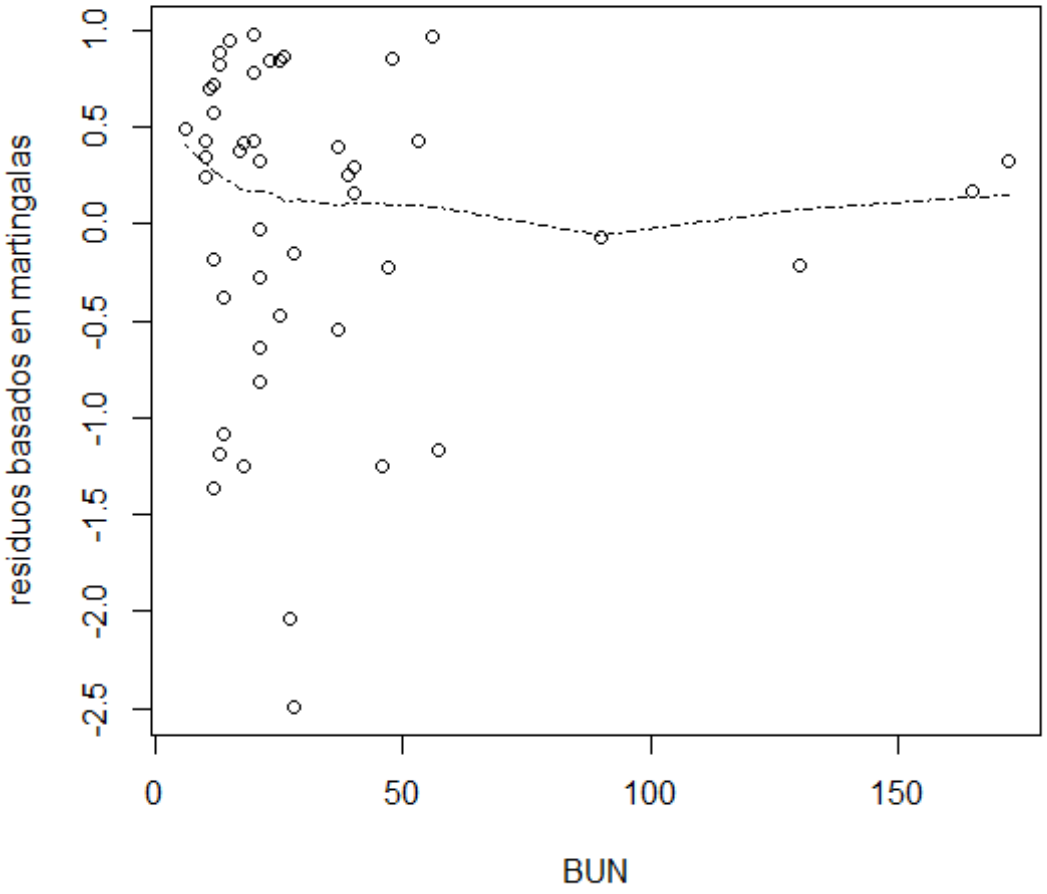

Figura 1: Gráfico de los residuos basados en martingalas vs BUN.

En ambos gráficos, Figura 1 y Figura 2, podemos ver que el suavizado da como resultado aproximadamente una recta (Dado que el modelo de Cox trabaja con aproximaciones y estimaciones se ha de ser flexible a la hora de interpretar los gráficos que vayamos obteniendo.), por lo tanto ninguna de las dos covariantes necesita una transformación.

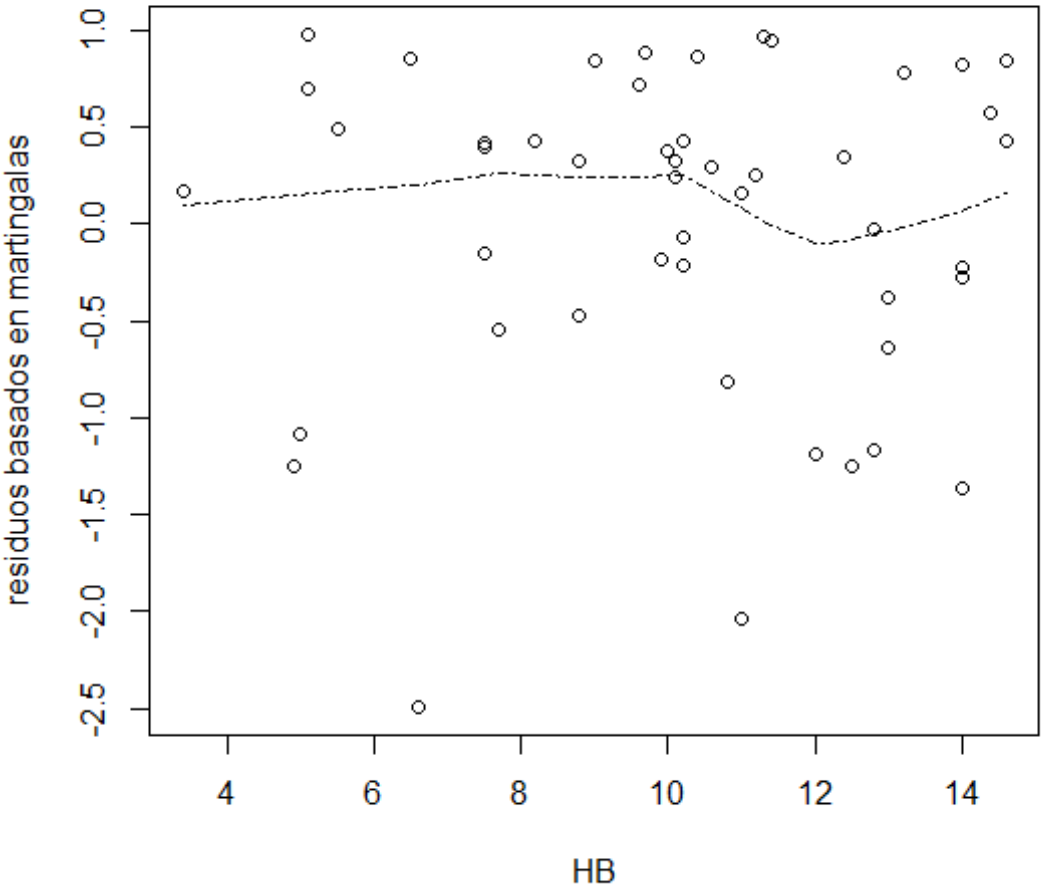

Figura 2: Gráfico de los residuos basados en martingalas vs HB.

El siguiente paso es comprobar la hipótesis de proporcionalidad utilizando los residuos de Schoenfeld. Mediante la función `cox.zph(dadesBUNHBBJP)` obtenemos unos *p-values*.

	p
BUN	0.662
HB	0.975
BJP	0.752
Global	0.968

Cuadro 9: Comprobación basada en los residuos de Schoenfeld.

Dado que todos los *p-values* son mayores que 0.05 no rechazamos la hipótesis de proporcionalidad. Otro método para comprobar que la hipótesis se cumple es mediante el gráfico de la función `cox.zph`. Utilizando

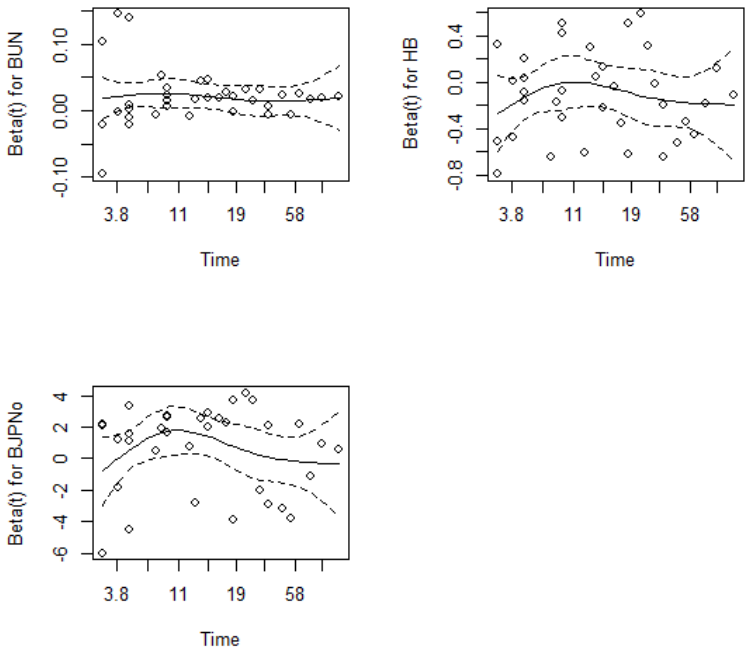

Figura 3: Residuos de Schoenfeld para las covariantes BUN, HB y BJP.

```
plot(cox.zph(dadesBUNHBBJP))
```

obtenemos la Figura 3.

Si la línea del estimador β a lo largo del tiempo es, aproximadamente, una recta entonces las covariantes cumplen la hipótesis de proporcionalidad. Podemos ver que en los tres gráficos de la Figura 3 se cumple. Por lo tanto, viendo que ambas pruebas indican que la hipótesis se cumple podemos afirmar que el modelo es razonable.

A continuación utilizaremos los residuos basados en la deviance para ver la validez del modelo para la predicción. Utilizando la formula

```
plot(resid(dadesBUNHBBJP, type="deviance"))
```

obtenemos la Figura 4.

Debido al limitado número de individuos la Figura 4 presenta un patrón poco definido. Aun así podemos ver que aunque los valores están bastante dispersos no hay grandes outliers, tan solo vemos dos individuos por encima de 2.

Por último debemos estudiar la influencia de cada individuo en el ajuste del modelo. Utilizaremos los residuos basados en el *score* como vimos en el apartado 4.4.3. Mediante la función

```
score $<-$ resid(dadesBUNHBBJP, type="dfbeta")
```

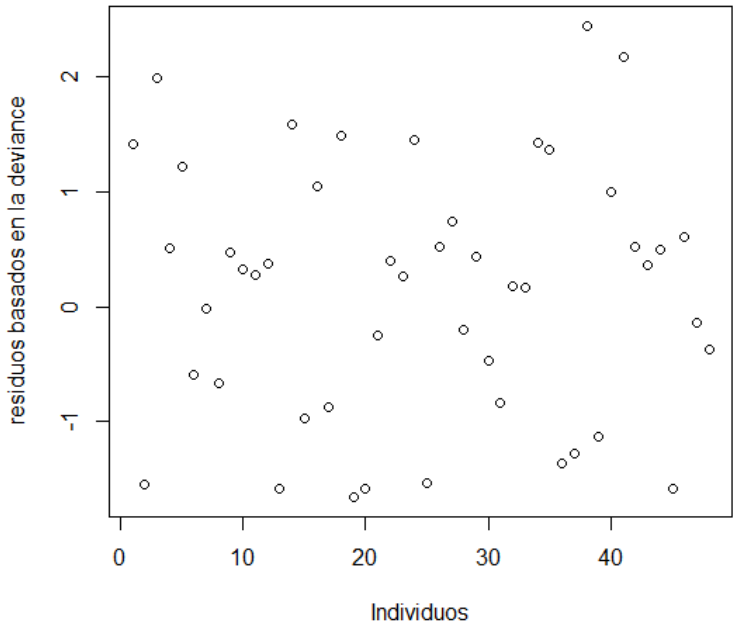

Figura 4: Residuos basados en la *deviance*.

obtendremos la matriz de residuos. Es muy importante recordar que está función devuelve una matriz y por lo tanto necesitaremos separar cada vector de residuos según a la covariante que corresponda. Si graficamos cada vector de residuos versus los individuos obtenemos la Figura 5.

Podemos ver de nuevo que no hay grandes outliers, vemos dos casos en los que el *score* está mucho más alejado que el resto para el BUN, tres para el HB y tres para el BJP. Por lo tanto es una buena suposición que nuestro modelo es razonablemente bueno.

Para terminar podemos estimar la función de supervivencia mediante el método de Kalbleisch y Prentice utilizando la función

```
survfit(dadesBUNHBBJP, type="kalbfleisch-prentice")
```

y graficandola. Si hacemos esto obtenemos la Figura 6.

5.4. Conclusiones del ajuste del modelo

Para terminar comentemos la información que podemos extraer de la ecuación 5.1.. El modelo nos dice que las covariantes que afectaran al riesgo de morir por mieloma múltiple son los niveles de urea en sangre, la cantidad de hemoglobina y la presencia de la proteína Bence-Jones. Por lo tanto en ua mujer y un hombre de distintas edades (recordemos que nuestro modelo solo es válido para edades entre

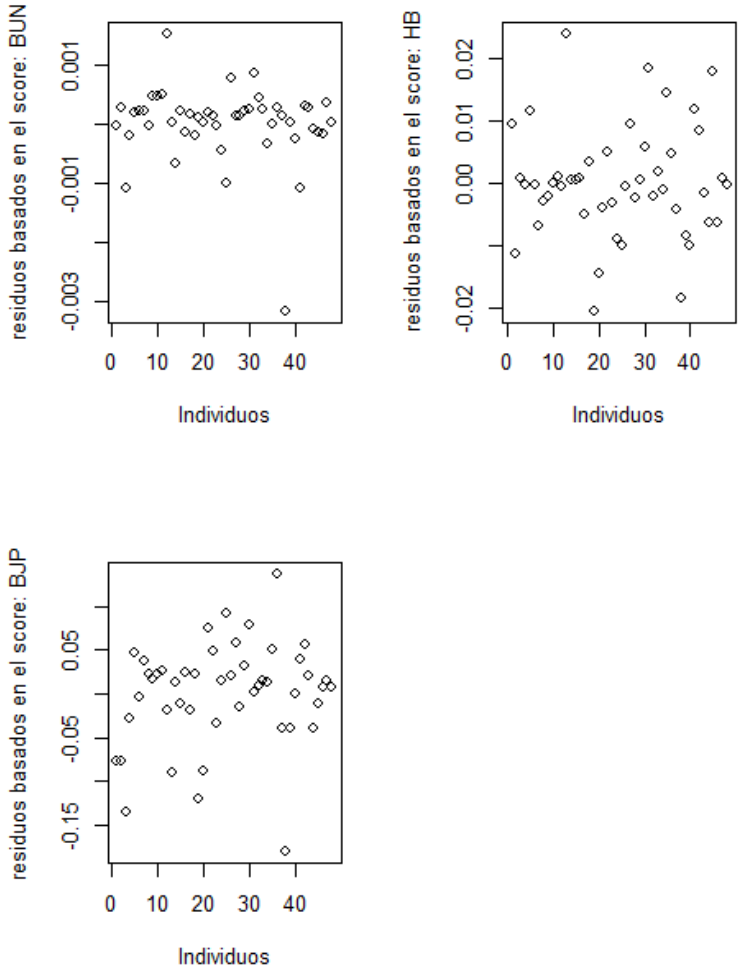

Figura 5: Residuos basados en los *score*.

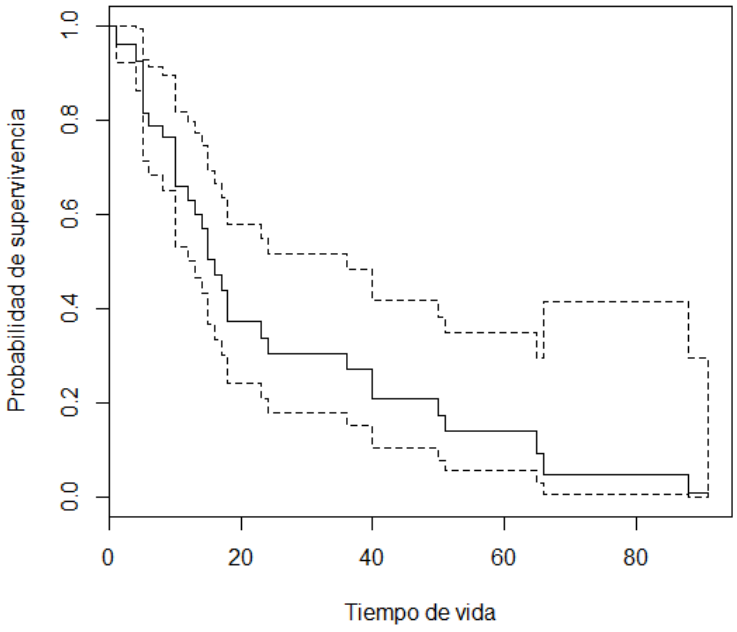

Figura 6: Gráfico de la probabilidad de supervivencia.

los 50 y 80 años) con los mismos valores de BUN, HC y BJ tendrán el mismo riesgo de morir por mieloma.

Ademas podemos ver que los coeficientes de BUN y BJP son positivos, por lo que un aumento del BUN o la presencia de la proteína Bence-Jones aumenta el riesgo de muerte. Veamos por ejemplo cual es el riesgo proporcional entre dos individuos con el mismo vector de covariantes salvo que el individuo 2 tiene 10 unidades más de BUN que el individuo 1.

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\lambda_0(t)e^{0{,}02043(Z_{BUN}+10)-0{,}11478Z_{HB}+0{,}61267Z_{BJP}}}{\lambda_0(t)e^{0{,}02043Z_{BUN}-0{,}11478Z_{HB}+0{,}61267Z_{BJP}}} = e^{0{,}02043\cdot 10} = 1{,}226. \quad (5.2)$$

Vemos que el riesgo de muerte del segundo individuo es 1,226 (22,6 %) superior.

Si ahora comparamos dos individuos con el mismo vector de covariantes salvo que uno tiene la proteína BJ (individuo 2) y otro no (individuo 1) veamos cual es el riesgo comparado.

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\lambda_0(t)e^{0{,}02043(Z_{BUN})-0{,}11478Z_{HB}+0{,}61267Z_{BJP1}}}{\lambda_0(t)e^{0{,}02043Z_{BUN}-0{,}11478Z_{HB}+0{,}61267Z_{BJP2}}} = \frac{e^{0{,}61267\cdot 1}}{e^{0{,}61267\cdot 0}} = e^{0{,}61267} = 1{,}845. \quad (5.3)$$

Podemos ver que la presencia de la proteína BJ aumenta el riesgo de morir por mieloma múltiple 1,845 (84,5 %) veces.

Por otro lado tenemos el coeficiente de HB que es negativo, por lo tanto si la hemoglobina aumenta, el riesgo de morir disminuye. Veamoslo con un ejemplo

numérico, dos individuos con el mismo vector de covariantes excepto el individuo 2 tiene $Z_{HB2} = Z_{HB} + 5$.

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\lambda_0(t)e^{0{,}02043(Z_{BUN})-0{,}11478(Z_{HB}+5)+0{,}61267Z_{BJP}}}{\lambda_0(t)e^{0{,}02043Z_{BUN}-0{,}11478Z_{HB}+0{,}61267Z_{BJP}}} = e^{-0{,}11478\cdot 5} = 0{,}891. \quad (5.4)$$

Un aumento de 5 unidades en el HB disminuye el riesgo de muerte en 1,21 (21 %).

Podemos concluir que un paciente sin la presencia de la proteína Bence-Jones, con unos niveles bajos de urea en sangre y con un numero elevado de hemoglobina tendrá menor riesgo de morir por mieloma múltiple.

6. Apéndice

Comandos utilizados en R para obtener los resultados del Capítulo 4.

```
myeloma
dadesage <- coxph(formula = Surv(time, status) ~ age , data = myeloma)
dadesage
anova(dadesage)
dades <- coxph(formula = Surv(time, status) ~ age + sex + BUN + CA + HB +  PC +
BJP, data = myeloma)
anova(dades)
AIC(dades)
AIC(dadesage)
dadessex <- coxph(formula = Surv(time, status) ~ sex , data = myeloma)
AIC(dadessex)
dadesBUN <- coxph(formula = Surv(time, status) ~ BUN , data = myeloma)
AIC(dadesBUN)
dadesCA <- coxph(formula = Surv(time, status) ~ CA , data = myeloma)
dadesHB <- coxph(formula = Surv(time, status) ~ HB , data = myeloma)
dadesPC <- coxph(formula = Surv(time, status) ~ PC , data = myeloma)
dadesBJP <- coxph(formula = Surv(time, status) ~ BJP , data = myeloma)

dadesage
dadessex
dadesBUN
dadesCA
dadesHB
dadesPC
dadesBJP

AIC(dadesCA)
AIC(dadesHB)
AIC(dadesPC)
AIC(dadesBJP)
dadesBUNage <- coxph(formula = Surv(time, status) ~ BUN + age ,
 data = myeloma)
dadesBUNsex <- coxph(formula = Surv(time, status) ~ BUN + sex ,
 data = myeloma)
dadesBUNCA <- coxph(formula = Surv(time, status) ~ BUN + CA ,
 data = myeloma)
dadesBUNHB <- coxph(formula = Surv(time, status) ~ BUN + HB ,
 data = myeloma)
dadesBUNPC <- coxph(formula = Surv(time, status) ~ BUN + PC
, data = myeloma)
dadesBUNBJP <- coxph(formula = Surv(time, status) ~ BUN + BJP ,
 data = myeloma)
```

```
AIC(dadesBUNage)
AIC(dadesBUNsex)
AIC(dadesBUNCA)
AIC(dadesBUNHB)
AIC(dadesBUNPC)
AIC(dadesBUNBJP)

dadesBUNHBage <- coxph(formula = Surv(time, status) ~ BUN + HB + age ,
 data = myeloma)
dadesBUNHBsex <- coxph(formula = Surv(time, status) ~ BUN + HB + sex ,
 data = myeloma)
dadesBUNHBCA <- coxph(formula = Surv(time, status) ~ BUN + HB + CA ,
 data = myeloma)
dadesBUNHBPC <- coxph(formula = Surv(time, status) ~ BUN + HB + PC ,
 data = myeloma)
dadesBUNHBBJP <- coxph(formula = Surv(time, status) ~ BUN + HB + BJP ,
 data = myeloma)
AIC(dadesBUNHBage)
AIC(dadesBUNHBsex)
AIC(dadesBUNHBCA)
AIC(dadesBUNHBPC)
AIC(dadesBUNHBBJP)

dadesBUNHBBJPage <- coxph(formula = Surv(time, status) ~ BUN + HB + BJP + age ,
 data = myeloma)
dadesBUNHBBJPsex <- coxph(formula = Surv(time, status) ~ BUN + HB + BJP  + sex ,
 data = myeloma)
dadesBUNHBBJPCA <- coxph(formula = Surv(time, status) ~ BUN + HB + BJP  + CA ,
 data = myeloma)
dadesBUNHBBJPPC <- coxph(formula = Surv(time, status) ~ BUN + HB + BJP  + PC ,
 data = myeloma)
AIC(dadesBUNHBBJPage)
AIC(dadesBUNHBBJPsex)
AIC(dadesBUNHBBJPCA)
AIC(dadesBUNHBBJPPC)

dadesBUNHBBJP
anova(dadesBUNHBBJP)
dades

martingalas <- resid(dadesBUNHBBJP, type="martingale")
plot(myeloma$age, martingalas, xlab='age', ylab='residuos basados en martingalas')
lines(lowess(myeloma$age, martingalas), lty=10)
plot(myeloma$BUN, martingalas, xlab='BUN', ylab='residuos basados en martingalas')
lines(lowess(myeloma$BUN, martingalas), lty=10)
plot(myeloma$CA, martingalas, xlab='CA', ylab='residuos basados en martingalas')
```

```
lines(lowess(myeloma$CA, martingalas), lty=10)
plot(myeloma$HB, martingalas, xlab='HB', ylab='residuos basados en martingalas')
lines(lowess(myeloma$HB, martingalas), lty=10)
plot(myeloma$PC, martingalas)
lines(lowess(myeloma$PC, martingalas), lty=10)

cox.zph(dadesBUNHBBJP)
par(mfrow=c(2,2))
plot(cox.zph(dadesBUNHBBJP))
cox.zph(dadesBUNHBBJP)

deviance <- resid(dadesBUNHBBJP, type="deviance")
plot(deviance, xlab='Individuos', ylab='residuos basados en la deviance')

score <- resid(dadesBUNHBBJP, type="dfbeta")
plot(score, myeloma$BJP)
score
score[1:48]
score[49:96]
score[97:144]
par(mfrow=c(2,2))
plot(score[1:48],xlab='Individuos', ylab='residuos basados en el score: BUN')
plot(score[49:96],xlab='Individuos', ylab='residuos basados en el score: HB')
plot(score[97:144],xlab='Individuos', ylab='residuos basados en el score: BJP')

help(survfit.coxph)
fit<-survfit(formula = Surv(time, status) ~ BUN + HB + BJP , data = myeloma)
survfit(dadesBUNHBBJP, type="kalbfleisch-prentice")
plot(survfit(dadesBUNHBBJP, type="kalbfleisch-prentice"))
plot(survfit(dadesBUNHBBJP, type="kalbfleisch-prentice"), xlab='Tiempo de vida',
 ylab='Probabilidad de supervivencia')
```

Referencias

[1] Gómez, Guadalupe; Julià, Olga; Langohr, Klaus: *Análisis de Supervivencia*. Bellaterra: Publicacions de la Universitat Politècnica de Barcelona, 2011.

[2] Terry M. Therneau, Patricia M. Grambsch: *Modeling Survival Data. Extending the Cox Model*. Springer Science+Business Media, LLC. Nueva York, 2000.

[3] John D. Kalbfleisch, Ross L. Prentice: The statical analysis of failure time data. John Wiley & Sons, Inc. 1980.