UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

MODELOS DE RIESGO DE SUPERVIVENCIA CON EFECTOS ADITIVOS Y MULTIPLICATIVOS

TESIS

QUE PARA OBTENER EL TíTULO DE:

Actuario

PRESENTA:

José de Jesús Velázquez Hernández

DIRECTORA DE TESIS:

Dra. Guillermina Eslava Gómez

Ciudad Universitaria, Cd. Mx., 2019

Datos de jurado

1. Datos del alumno
Velázquez
Hernández
José de Jesús
57 42 32 17
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
311134852
2. Datos del tutor
Dra.
Guillermina
Eslava
Gómez
3. Datos del sinodal 1
Mat.
Margarita Elvira
Chávez
Cano
4. Datos del sinodal 2
Dra.
Sofía
Villers
Gómez
5. Datos del sinodal 3
M. en C.
David Chaffrey
Moreno
Fernández
6. Datos del sinodal 4
M. en C.
Graciela
Martínez
Sánchez
7. Datos del trabajo escrito
Modelos de riesgo de supervivencia
con efectos aditivos y multiplicativos
129 p
2019

Agradecimientos

A mis padres, por su constante apoyo y amor.

A la Dra. Guillermina Eslava Gómez, por sus enseñanzas, su disposición y su perseverancia en la dirección de este trabajo.

A mis sinodales, por sus sugerencias y aportaciones; y a todo el cuerpo docente de la Facultad de Ciencias que me inspiró durante mi formación con su entrega y dedicación.

A mis amigos y familiares, por ser parte de mis mejores recuerdos, por motivarme y creer en mí. Ellos saben quiénes son.

A todos quienes me apoyaron, les estoy sinceramente agradecido.

Resumen

El presente trabajo tiene como objetivo ejemplificar los modelos de riesgos proporcionales, aditivo y multiplicativo-aditivo con un enfoque a los procesos de conteo. Como caso típico, se analizaron dos bases de datos extraídas de la bibliografía. La primera consiste en 877 observaciones de 23 variables correspondientes a un estudio de pacientes infectadas de clamidia o gonorrea. Se describe el riesgo de reinfección mediante un ajuste de modelos de regresión proporcionales, aditivo y multiplicativo-aditivo. Los efectos descritos en los modelos ajustados son constantes, pues se rechaza la hipótesis de coeficientes variables a lo largo del tiempo de observación para cada covariable con un nivel de confianza del 95 %. Estos modelos permiten asignar un índice de pronóstico para cada paciente y los terciles de esta medida determinan 3 grupos de riesgo: bajo, mediano y alto. Para evaluar los modelos ajustados, se realiza un análisis de residuales de martingalas, de devianza y de score para evaluar la bondad de cada ajuste.

La segunda base describe el tiempo de fallecimiento de un grupo de 418 pacientes con cirrosis biliar primaria mediante 19 variables. Al ajustar los modelos de regresión, se rechaza la hipótesis de coeficientes constantes para 2 de 5 covariables en cada modelo. Por ello, se hace un ajuste de modelos de riesgos proporcionales y aditivos semi paramétricos. Una vez determinadas cuáles variables tienen un mejor ajuste en el modelo proporcional y cuáles en el modelo aditivo, se ajusta el modelo multiplicativo-aditivo. Finalmente, y como en el primer análisis, se comprueba la exactitud de cada ajuste a través del análisis de residuales de martingala, de score y de devianza.

Índice general

Introducción

Desde hace medio siglo, se han desarrollado métodos estadísticos que han permitido analizar el comportamiento del tiempo que ocurre hasta cierto evento de interés en datos censurados y el efecto que tienen otras variables en él. Esto se ha aplicado en distintos contextos donde se tienen datos longitudinales. No obstante, se ha desarrollado la teoría de Procesos de conteo que ha complementado el análisis de supervivencia, de tal modo que permite probar hipótesis con mayor precisión y explicar el riesgo a través de variables con más detalle.

En los primeros dos capítulos se parte de algunos resultados del análisis de supervivencia y un preámbulo de la teoría de los procesos de conteo. Posteriormente, se presentan los modelos de riesgos proporcionales, aditivos y multiplicativo-aditivo. Se describe *grosso modo* la estimación de los parámetros de los modelos y sus propiedades asintóticas, las pruebas de hipótesis más importantes y se introduce al análisis de residuales, con el cuál se aplica la teoría de procesos de conteo. Con el fin de ilustrar dicha metodología, se presentan en el tercer capítulo, estudios de bases de datos obtenidas de la literatura. Se analiza primeramente la base *std* de la paquetería *KMsurv* en R, la cual considera un grupo de 877 pacientes femeninas que en un principio habían padecido clamidia o gonorrea (o ambas). Siendo así, se tiene una base de datos con variables como la edad, años de escolaridad, tipo de infección adquirida, número de parejas y uso del condón. Adicionalmente, se tiene información acerca de ciertas señales observadas durante el diagnóstico médico de la paciente como el dolor abdominal, disuria, sarpullido, lesiones, brote de reinfección y el tiempo hasta que se presentó. El objetivo del análisis es identificar cuáles de estas variables tienen un efecto más importante en el tiempo de recurrencia de las enfermedades.

Por otro lado, la base *pbc*, obtenida de la paquetería *survival* en R, contiene las variables del tiempo hasta la muerte por cirrosis, de transplante o de fin del estudio, el status médico del paciente, tratamiento administrado, edad y sexo. Algunas otras variables de señales obtenidas en el diagnóstico, son la presencia de ascitis, hepatomegalia, malformaciones venales, de edema; el nivel de bilirubina, colesterol, albumina, triglicéridos y el tiempo que toma en coagular la sangre. Durante el análisis se obtienen hallazgos acerca de la forma en la que las variables anteriores influyen en el tiempo de fallecimiento.

Capítulo 1

Procesos de conteo en el análisis de Supervivencia

En estudios de fenómenos relacionados con la supervivencia de alguna población (sean pacientes, poblaciones de animales, entidades financieras, aparatos, etc..), es de gran interés observar el tiempo de supervivencia de los elementos de la muestra con el fin de obtener conclusiones de la población sustentadas en la observación de los individuos. Sin embargo, pueden presentarse problemas debido a la dificultad del seguimiento del tiempo de supervivencia de algunos individuos. Los estudios con las características anteriores pueden ser desarrollados mediante el Análisis de Supervivencia, cuyo objetivo es ajustar un modelo estadístico que permita identificar los factores de riesgo que tienen un impacto en la población de tal manera que se obtengan conclusiones sobre el tiempo de supervivencia de la población.

Tipos de falla y censura

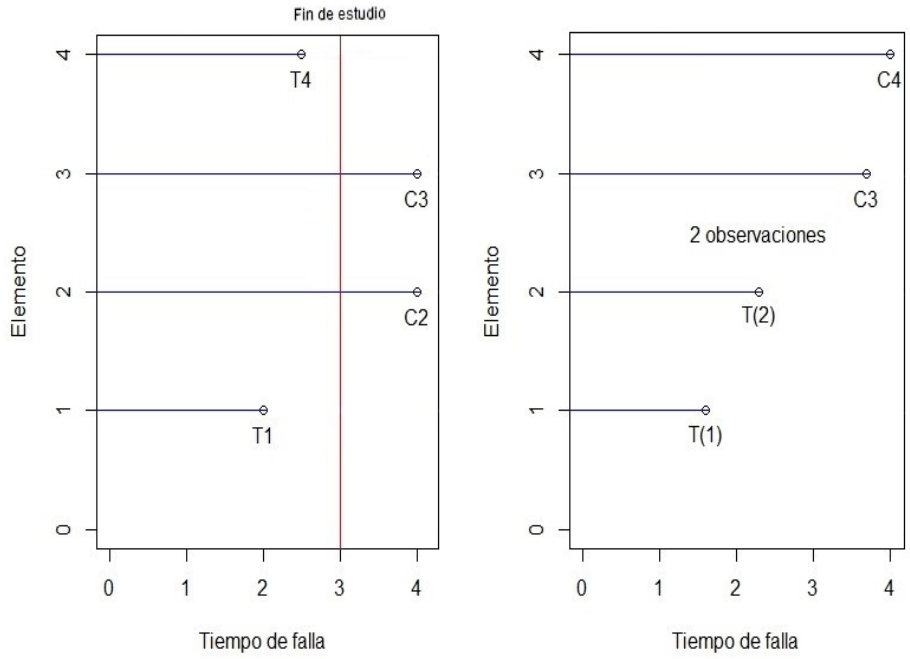
Lawless (2003 pág. 53) define el tiempo de falla como $T_i = min\{X_i, C_i\}$, el mínimo entre el tiempo de vida X_i (el tiempo en el que el individuo no presenta el evento de interés) y el tiempo de censura C_i (el tiempo en el que se termina el estudio o que el individuo sale del grupo de estudio por alguna causa ajena al evento de interés). Ambas variables X_i y C_i toman valores en el intervalo $[0, \infty)$ y por lo tanto T también.
Se debe prestar atención a la definición del tiempo de censura, pues tiene especial efecto en el análisis de supervivencia. *Cox & Oakes* (1984, pág. 4) afirman que, como el tiempo de supervivencia, la censura es un evento durante el período de observación con el cual termina el seguimiento de la supervivencia del individuo, si es que el evento de interés no ocurre hasta ese punto.

La censura puede presentarse por diversas causas, como la salida voluntaria de algún individuo en el estudio o porque el investigador decide terminar el período de observación debido a la falta de recursos o por propósitos de la investigación. Por ello se clasifican 2 tipos de censura:

- **Censura tipo I**: Ocurre cuando se decide terminar el período de observación de los individuos hasta cierto tiempo determinado. Si los individuos no presentaron un tiempo de falla hasta entonces, se dice que presentan censura de tipo I.
- **Censura tipo II**: Se presenta una vez que el investigador decide que han ocurrido un número determinado de fallas k de los n individuos. Así el tiempo de censura C se convierte en $T_{(k)}$ el k-ésimo estadístico de orden de la muestra.

La **fig. 1.1** muestra un ejemplo de estos tipos de censura. Intrínsecamente se supondrá a lo largo de este trabajo que los tiempos de censura son independientes para todos los individuos, ya que se tienen casos en los que las fallas de los individuos alteran la intensidad o el riesgo de otros individuos en la población. La forma en la que se incluyen estos tiempo en el análisis es considerando el tipo I de censura.

Figura 1.1: Ejemplo de tipo de censura I (izquierda) para un tiempo de estudio de 3 unidades y tipo de censura II (derecha) para las primeras 2 observaciones.



Funciones involucradas en el análisis de supervivencia

La función de densidad de probabilidad se define del siguiente modo para una variable aleatoria discreta T de tiempo de falla:

$$f(t) = P[T = t]$$

Dependiendo de la forma de modelar la variable aleatoria del tiempo de falla T, si es continua o discreta, se interpretarán teóricamente las funciones de probabilidad de un modo distinto. Para este trabajo se considerará la variable del tiempo de falla T como continua. Las definiciones, extraídas de *Lawless* (2003, pág. 8) se resumen a continuación.

Función de supervivencia

Es la probabilidad de que un individuo que sobreviva en el tiempo t, y satisface ser monótona decreciente:

$$S(t) = P[T \geq t] = \int_t^\infty f(x)dx$$

Función de riesgo

Especifica la tasa instantánea de riesgo al tiempo t, dado que un individuo sobrevive al tiempo t, de modo que $h(t)\Delta t$ es aproximadamente la probabilidad de falla en el intervalo infinitesimal $[t, t + \Delta t)$, dado el tiempo de supervivencia t. Se define como:

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

Además, usando la definición de probabilidad condicional y sustituyendo la expresión de la función de supervivencia, se obtiene:

$$h(t) = \frac{f(t)}{S(t)} = \left(-\frac{d}{dt}S(t)\right)\frac{1}{S(t)} = -\frac{d}{dt}\log S(t) \tag{1.0.1}$$

En el contexto de martingalas, la función de riesgo se conoce también como función de intensidad, por su relación con los procesos estocásticos.

Función de riesgo acumulada

La función de riesgo acumulada se define como

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

Esta función es importante para poder observar la frecuencia con la que ocurren las fallas a lo largo del tiempo y será de gran utilidad para analizar los residuos para la evaluación del ajuste de los modelos. De la expresión **1.0.1** se obtiene:

$$H(t) = -\log S(t)$$

Estas funciones describen la distribución del tiempo de falla y son de gran utilidad para realizar el análisis. A continuación, se presenta la relación entre estas funciones.

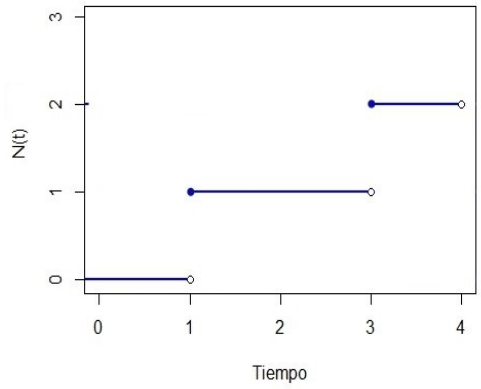
Cuadro 1.1: Relación entre funciones de probabilidad. Cada casilla representa la función de la columna expresada en la columna en términos de la función del renglón correspondiente.

	$f(t)$	$S(t)$	$h(t)$	$H(t)$
$f(t)$		$\int_t^\infty f(x)dx$	$\frac{f(t)}{\int_t^\infty f(x)dx}$	$\int_0^t \frac{f(x)}{\int_x^\infty f(u)du}dx$
$S(t)$	$-\frac{d}{dt}S(t)$		$-\frac{d}{dt}\log S(t)$	$-\log S(t)$
$h(t)$	$h(t)exp(-\int_0^t h(x)dx)$	$exp(-\int_0^t h(x)dx)$		$\int_0^t h(x)dx$
$H(t)$	$-\frac{d}{dt}exp(-H(t))$	$exp(-H(t))$	$\frac{d}{dt}H(t)$	

Procesos de conteo

Las funciones de probabilidad definidas anteriormente sirven para caracterizar la variable de tiempo de falla T por completo, siempre que se suponga una distribución paramétrica para T. Sin embargo, observar otros enfoques en el análisis de supervivencia puede aportar resultados interesantes, tal como los procesos estocásticos, que en vez de enfocarse en la variable T que rige los tiempos de falla, describen un proceso de conteo para representar las fallas de la muestra.

Figura 1.2: Ejemplo de proceso de conteo $N(t)$.



Para describir este enfoque, *Aalen* (2008, pág. 53) define un *Proceso de conteo* $N = \{N(t) : t \in [0, \tau]\}$ como un proceso continuo por la derecha con saltos unitarios, en los tiempos en los que se presenta algún evento, y constante entre ellos, tal como se muestra en la **fig. 1.2**. Un ejemplo de dicho proceso, es el *Proceso Poisson*, que modela el conteo de eventos entre los cuales el tiempo se distribuye de forma exponencial. Este proceso se describe mediante su función de intensidad λ, de la cual se interpreta $\lambda(t)dt$ como la probabilidad de ocurrencia del evento de interés en un intervalo de tiempo infinitesimal.

Así, recordando la definición de la función de riesgo $h(t)$ del tiempo de falla T, se tiene una forma de expresar *la función de intensidad* $\lambda(t)$ para cada individuo i de acuerdo con *Aalen* (2008, pág. 70, exp. 3.1):

$$\lambda_i(t) = Y_i(t)h(t)$$

en donde se define la variable $Y_i(t)$ para cada individuo i de la siguiente manera:

$$Y_i(t) = \mathbb{1}_{\{T_i \geq t\}} = \begin{cases} 1 & si \quad T_i \geq t \\ 0 & si \quad T_i < t \end{cases} \tag{1.0.2}$$

la cual es una función que indica si al individuo i aún no le ocurre la falla o la censura en el tiempo t. Es decir, indica la exposición al riesgo. Dentro del intervalo de tiempo en el que el individuo está expuesto, la función de riesgo de la distribución del tiempo de falla es la función de intensidad del proceso de conteo para el individuo i.

Dado lo anterior, se define la función de intensidad acumulada con la siguiente expresión:

$$\Lambda(t) = \int_0^t \lambda(s)ds \tag{1.0.3}$$

Los eventos a contar que serán de interés son precisamente las fallas que ocurran hasta el tiempo t. En el contexto de Análisis de supervivencia, el proceso de conteo $N(t)$ sólo suele tomar a lo más el valor de 1, pues corresponde únicamente al número de fallas que presenta el i-ésimo elemento del estudio y generalmente se conoce la falla como un *estado absorbente*, aunque dependiendo del fenómeno que se estudie, puede tomar valores en todo el conjunto de los números naturales, en el caso en que se trate de un estado recurrente. Como en el ejemplo de la **fig. 1.2**.

Considerando lo anterior, la idea de este enfoque del análisis de supervivencia, es explicar el proceso de conteo $N(t)$ mediante la forma *Variable respuesta=Señal+Ruido*. Antes de poder justificar este modelo, y para abordar el tema adecuadamente, es necesario definir dos clases de procesos estocásticos.
Aalen (2008, pág. 48) define una martingala como un proceso $M = \{M(t) : t \in [0, \tau]\}$ adaptado a la filtración $\mathscr{F}_t$ que cumple la siguiente propiedad:

$$E\left[M(t)|\mathscr{F}_s\right] = M(s)\ \forall s < t.$$

Si sólo se cumple que $E\left[M(t)|\mathscr{F}_s\right] \leq M(s)\ \forall s < t$, se dice que $M(t)$ es una submartingala. De lo anterior, se puede concluir que $N(t)$ es una submartingala por ser un proceso no decreciente. Y de este modo, la *Descomposición de Doob-Meyer* formula que cualquier submartingala (en este caso $N(t)$) puede ser descompuesta de la siguiente manera:

$$N(t) = \Lambda(t) + M(t) \tag{1.0.4}$$

Donde $M(t)$ es una martingala y $\Lambda(t)$ se conoce *compensador* del proceso $N(t)$. Para entender el papel de $\Lambda(t)$ en la expresión **1.0.4**, *Therneau & Grambsch* (2000, pág. 5) definen intuitivamente

un *proceso predecible* $\Lambda(t)$, como aquél cuyo valor en el tiempo t se conoce en un tiempo infinitesimalmente menor $(t-)$, o bien, formalmente se define como un proceso que es adaptado a la filtración $\mathscr{F}_{t-}$.
Este proceso $\Lambda(t)$ será justamente aquél que representará el riesgo dentro de los modelos de regresión y aquél que se desea estimar. Más aún, juega un papel muy importante, pues es justamente el *compensador* de la *descomposición de Doob-Meyer* de la ecuación **1.0.4**. Es decir, $\Lambda(t)$ es el componente que aporta la tendencia creciente de $N(t)$. Sin este componente, $N(t)$ no sería ni submartingala ni proceso de conteo.

Y finalmente, hay que mencionar que bajo este enfoque se considera la siguiente notación matricial, donde $N(t) = (N_1(t), ..., N_n(t))$ representa cada elemento $i = 1, ..., n$ de la muestra. Análogamente se considera el vector $\Lambda(t) = (\Lambda_1(t), ..., \Lambda_n(t))$ y $M(t) = (M_1(t), ..., M_n(t))$.

Estimadores de Nelson-Aalen y Kaplan-Meier

Para poder ajustar los modelos, será necesario estimar la función de intensidad acumulada definida en **1.0.3**, y para construir tal estimador, se definen $N.(t) = \sum_{i=1}^{n} N_i(t)$ como el número total de fallas que se registran para la población y $Y.(t) = \sum_{i=1}^{n} Y_i(t)$ como el total de individuos expuestos hasta el tiempo t.

De acuerdo con la definición de la integral de Riemman-Stieltjes, el estimador de Nelson-Aalen se escribe como:

$$\hat{\Lambda}(t) = \int_0^t \frac{dN.(s)}{Y.(s)}$$

para el caso continuo. Para el caso discreto, se puede escribir como:

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{\Delta N.(t_i)}{Y.(t_i)}$$

Es importante aclarar que $dN.(t) = N.(t + dt) - N.(t)$, o bien, $\Delta N.(t) = N.(t + \Delta t) - N.(t)$ puede interpretarse como una variable aleatoria Bernoulli (pues en un intervalo muy pequeño de tiempo sólo puede ocurrir o no ocurrir un evento), cuyo parámetro es la probabilidad de ver un evento en un intervalo infinitesimal alrededor de t, es decir:

$$P(dN(t) = 1|\mathscr{F}_{t-}) = E(dN(t)|\mathscr{F}_{t-}) = E(d\Lambda(t) + dM(t)|\mathscr{F}_{t-})$$

$$= d\Lambda(t) + E(dM(t)|\mathscr{F}_{t-}) = \lambda(t)dt$$

Esta equivalencia se debe a que el proceso $\Lambda(t)$ es un proceso predecible (y por lo tanto $\mathscr{F}_{t-}$ medible) y que para las martingalas $M(t)$ se cumple que $E(dM(t)|\mathscr{F}_{t-}) = 0$.

La varianza de este estimador, que nos permitirá construir intervalos de confianza para la función de supervivencia, está dada por la siguiente expresión:

$$Var(\hat{\Lambda}(t)) = \int_0^t \frac{dN.(s)}{Y.^2(s)}$$

o bien, para el caso discreto:

$$Var(\hat{\Lambda}(t)) = \sum_{i:t_i \leq t} \frac{\Delta N.(t_i)}{Y.^2(t_i)}$$

En este caso, la función de supervivencia se puede expresar, dada la relación que tiene con la función de riesgo acumulado expresada en el **cuadro 1.1**, de la siguiente manera:

$$\hat{S}_{NA}(t) = \exp\{-\hat{\Lambda}(t)\} = \exp\left\{-\sum_{i:t_i \leq t} \frac{\Delta N.(t_i)}{Y.(t_i)}\right\} = \prod_{i:t_i \leq t} \exp\{-\Delta\hat{\Lambda}(t_i)\}$$

pues $\Delta\hat{\Lambda}(t_i) = \Delta N.(t_i)/Y.(t_i)$. Por otro lado, el estimador de Kaplan-Meier está dado por la siguiente expresión:

$$\hat{S}_{KM}(t) = \prod_{i:t_i \leq t} (1 - d\hat{\Lambda}(t_i))$$

Se puede observar la aproximación al estimador obtenido por Nelson-Aalen, pues $e^x \approx 1 - x$ para $x \to 0$ (que generalmente es el rango de valores que toma $\Delta\hat{\Lambda}(t_i)$). Con ello se muestra que ambos estimadores producen resultados muy similares.
Siendo así, y usando su relación con la función de supervivencia, el estimador obtenido por Kaplan-Meier para la función de riesgo acumulada quedaría expresado como:

$$\hat{\Lambda}_{KM}(t) = -\log(\hat{S}_{KM}(t))$$

Además, para construir intervalos de confianza, un estimador consistente para la varianza de $\hat{S}_{KM}(t)$ está dado por la fórmula de Greenwood, cuyo desarrollo es presentado, por ejemplo en *Cox & Oakes* (1984, pág. 50):

$$\widehat{Var}(\hat{S}_{KM}) = \hat{S}^2(t) \sum_{i:t_i \leq t} \{Y.(t_i)(Y.(t_i) - \Delta N.(t_i))\}^{-1} \Delta N.(t_i)$$

de lo cual se obtienen, para cada tiempo t, intervalos de la forma:

$$\hat{S}(t) \pm z_{1-\alpha/2} \left[\widehat{Var}(\hat{S}(t))\right]^{\frac{1}{2}}$$

donde $z_{1-\alpha/2}$ es el cuantil $1 - \alpha/2$ de una distribución normal estándar.

Prueba de *log rank*

A lo largo de los análisis realizados en el **capítulo 3**, será preciso contrastar las diferencias entre las funciones de supervivencia correspondientes a diferentes grupos o tratamientos de pacientes. La prueba de *log rank* se aplica generalmente para determinar la diferencia entre dos grupos, pero a continuación se describe el contraste del caso general para p grupos o tratamientos. Es decir, se desea contrastar la siguiente hipótesis:

$$H_0 : S_i(t) = S_j(t) \ \forall i, j = 1, ..., p \quad vs. \quad H_1 : S_i(t) \neq S_j(t) \text{ para algunos } i, j = 1, ..., p \tag{1.0.5}$$

Sea t_i el tiempo en el que se presenta la i-ésima de las n observaciones. Dichas observaciones contienen todos los grupos, los cuales no tienen que ser del mismo tamaño. El número de elementos de la muestra expuestos o en riesgo (que no han presentado el evento) se determina para el k-ésimo grupo y se denota como $Y_{\cdot k}(t)$ con $k = 1, ..., p$. Además, sea $\Delta N_{\cdot}(t_i) = N_{\cdot}(t_i) - N_{\cdot}(t_{i-1})$ el número de fallas que ocurren al tiempo t_i (empates en tiempo de falla de las observaciones). Con estas cantidades, se estima el número esperado de fallas en el grupo k al tiempo T_i, dado por:

$$E_k(T_i) = \frac{\Delta N_{\cdot}(t_i) \cdot Y_{\cdot k}(t_i)}{Y_{\cdot}(t_i)}$$

Este cálculo se realiza para cada tiempo de falla t_i y cada grupo observado en la muestra. Finalmente, se determina el número total de fallas observadas O_k y la suma de las fallas esperadas $E_k = \sum_{i=1}^{n} E_k(T_i)$ de cada uno de los grupos. Con ello se construye la siguiente estadística que se distribuye como $\chi^2_{(p-1)}$:

$$T = \sum_{i=1}^{p} \frac{(O_i - E_i)^2}{E_i}$$

La hipótesis nula **1.0.5** se rechaza si $T > \chi^2_{(p-1),1-\alpha}$, donde $\chi^2_{(p-1),1-\alpha}$ es el cuantil $1 - \alpha$ de una distribución $\chi^2_{(p-1)}$. Dado que la prueba puede no ser determinante para las curvas de $S(t)$ que se cruzan u oscilan en algunos grupos, se recomienda el apoyo gráfico.

Capítulo 2

Modelos de regresión

En este capítulo se describen los modelos de regresión con variables explicativas, cuyo objetivo es describir los efectos de dichas variables en el tiempo de falla de los datos de la muestra. Se presentan los modelos de riesgos proporcionales junto con algunas generalizaciones (modelos no paramétrico y semi paramétricos) con el fin de compararlos y así poder determinar cuál es el modelo más adecuado para cada caso. Se aborda el tema definiendo el modelo de riesgos proporcionales simple para después dar paso a los modelos aditivos y multiplicativo-aditivos como otras formas de modelar la función de intensidad $\lambda(t)$ del proceso de conteo $N(t)$, el cual es el principal objetivo.

De acuerdo con *Cox & Oakes* (1984, pág. 91, exp. 7.1), los modelos de riesgos proporcionales se definen para el tiempo t, dada una matriz $\boldsymbol{Z}_{n\times p} = (Z_1, ..., Z_n)^T$ con las variables explicativas para cada individuo de la muestra, de la siguiente manera:

$$h(t|\boldsymbol{Z}) = h_0(t)\alpha(\boldsymbol{Z};\beta) \tag{2.0.1}$$

donde $\alpha(\boldsymbol{Z};\beta)$ es el factor de proporcionalidad que aporta la información de las variables $\boldsymbol{Z}$ y $h_0(t)$ es la función de riesgo de base que representa la función de riesgo para un elemento de la población bajo características estándares, es decir, siendo $\boldsymbol{Z} = 0$. Por ello, se requiere $\alpha(\mathbf{0};\beta) = 1$. Según *Cox & Oakes* (1984, pág. 70), el modelo anterior se basa en la familia de distribuciones de Lehmann, las cuales se expresan en términos de las funciones de supervivencia y densidad basales $S_0(t)$ y $f_0(t)$:

$$S(t|\boldsymbol{Z}) = S_0(t)^{\alpha(\boldsymbol{Z})}; \quad f(t|\boldsymbol{Z}) = \alpha(\boldsymbol{Z})S_0(t)^{\alpha(\boldsymbol{Z})-1}f_0(t)$$

Las expresiones anteriores se obtienen de la función de riesgo **2.0.1** y el **cuadro 1.1** del capítulo anterior. Las variables que se consideran dentro de las bases de datos a analizar por el modelo elegido, serán incorporadas en la matriz $\boldsymbol{Z}$ de dimensiones $n \times p$, que contendrá la información de los n individuos y las p variables. Generalmente, y en los ejemplos presentados en el **capítulo 4**, las variables no dependen del tiempo para cada individuo. Aunque se elige la nota-

ción $\boldsymbol{Z}(t)$ para indicar que la expresión es válida para las variables que sí tienen esa característica.

De tal modo, **2.0.1** es una forma de representar la función de riesgo $h(t|\boldsymbol{Z})$ de tiempo de falla en proporción a otra función de riesgo para una población con riesgos estándares $h_0(t)$, considerando ciertos factores de riesgos $\boldsymbol{Z}$. La función de riesgos $h(t|\boldsymbol{Z})$ puede escribirse sólo como $h(t)$ y el factor de proporcionalidad, dado por una función $\alpha(\boldsymbol{Z};\beta)$ [1], puede ser cualquier función que cumpla $\alpha(\mathbf{0};\beta) = 1$, ya que su interpretación dependerá de su forma. Algunos ejemplos de α son los siguientes:

- $\alpha(\boldsymbol{Z};\beta) = \log_2(1 + e^{\boldsymbol{Z}^T\beta})$
- $\alpha(\boldsymbol{Z};\beta) = 1 + \boldsymbol{Z}^T\beta$
- $\alpha(\boldsymbol{Z};\beta) = e^{\boldsymbol{Z}^T\beta}$

Si bien, la teoría y aplicaciones de cada una de las expresiones de α son interesantes, sólo se profundizará en la última expresión. La cual es definida en la siguiente sección.

2.1. Modelo de riesgos proporcionales

Este modelo ha sido ampliamente difundido gracias a autores como Cox. Consiste en definir el factor proporcional $alpha$ como $\alpha(\boldsymbol{Z};\beta) = e^{\boldsymbol{Z}^T\beta}$ de tal manera que se defina la función vector de intensidad como:

$$\lambda(t) = Y(t)\lambda_0(t)\exp(\boldsymbol{Z}^T(t)\beta) \tag{2.1.1}$$

Donde $\lambda_0(t) = Y(t)h_0(t)$, con $h_0(t)$ definida como en la expresión **2.0.1**. Se considera que, para cada individuo en el tiempo t, $\boldsymbol{Z}(t)$ es la matriz de observaciones para las p variables en cuestión y $\beta = (\beta_1, ..., \beta_p)$ es el respectivo vector p-dimensional para los coeficientes de las variables en el modelo. $Y(t)$ es la función indicadora para denotar el tiempo de observación de la población en riesgo para el tiempo t definida en **1.0.2**.

Esta versión simple del modelo de riesgos proporcionales, de acuerdo con autores como *Martinussen & Scheike* se conoce como modelo de riesgos proporcionales *paramétrico*, lo cuál no es usual en el resto de la literatura del análisis de Supervivencia. El razonamiento detrás de esta forma de referirse al modelo yace en que se pueden determinar los parámetros que conforman el vector β (cosa que no ocurre cuando el vector $\beta(t)$ depende del tiempo) y con ello se puede hacer predicción de otros valores. En el caso de que el vector de coeficientes $\beta(t) = (\beta_1(t), ..., \beta_p(t))$ dependa del tiempo, se dice que se trata de un *modelo no paramétrico* porque el modelo no se puede especificar con un vector de coeficientes fijo. En este caso se estiman los valores de $\beta(t)$ para cada tiempo de falla t durante el tiempo de observación.

[1] En adelante escrito como $\alpha(\boldsymbol{Z})$

Estimación e interpretación de los coeficientes del modelo

Para el caso del modelo paramétrico, los parámetros que lo determinan son el vector de coeficientes β que representa el efecto que tienen las variables en el modelo de **2.1.1**, los cuales se obtienen de la función de verosimilitud, la cual se construye de la siguiente manera descrita por *Cox & Oakes* (1984, pág 91), suponiendo que T es una variable aleatoria discreta.

Primeramente se observan los tiempos de falla ordenados de los n individuos $t_{(1)} < t_{(2)} < ... < t_{(n)}$ (se supone que no hay empates) y el j-ésimo conjunto de riesgo denotado como $R(j) = \{l = 1, ..., n : t_l \geq t_{(j)}\}$ para los tiempos de falla superiores a $t_{(j)}$. Seguidamente, se considera la probabilidad de que el i-ésimo individuo con el vector de covariables Z_i muera al tiempo t_i dado que pertenece al grupo de riesgo $R(j)$. Esto se denotará como $p_i(R(j))$, y siendo T_i el tiempo de falla:

$$p_i(R(j)) = \frac{P[T_i = t_i, \boldsymbol{Z} = Z_i]}{P[i \in R(j)]} = \frac{h(t_i|Z_i)}{\sum_{k \in R(j)} h(t_i|Z_k)} = \frac{Y_i(t)\exp(\boldsymbol{Z}_i^T(t)\beta)}{\sum_{k \in R(j)} Y_k(t)\exp(\boldsymbol{Z}_k^T(t)\beta)} \quad (2.1.2)$$

La expresión **2.1.2** corresponde a la i-ésima contribución de la función de verosimilitud del modelo donde las probabilidades son de eventos independientes, de tal modo que la función de verosimilitud se obtiene con la siguiente expresión:

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} p_i = \prod_{i=1}^{n} \frac{\exp(\boldsymbol{Z}_i^T(t)\beta)}{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{Z}_j^T(t)\beta)}$$

Además, tomando en cuenta los empates o elementos que fallan al mismo tiempo, se puede escribir la expresión anterior tal como la formulan *Fleming & Harrington* (1991, pág. 142, exp. 3.7) usando la notación matricial y asociando las correspondientes entradas de los vectores $N(t)$, $Y(t)$ y el correspondiente vector de la matriz $\boldsymbol{Z}(t)$:

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{Z}_i^T(t)\beta)}{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{Z}_j^T(t)\beta)} \right)^{\Delta N.(t_i)} \quad (2.1.3)$$

Donde $\Delta N.(t_i) = N.(t_i) - N.(t_{i-1})$ representa el salto en el proceso por los individuos que fallan en el mismo intervalo de tiempo, esto para considerar los empates. Para la estimación del vector β se aplica la función logaritmo. Es decir, para los tiempos de falla $t_1, ..., t_n$:

$$\log(\mathscr{L}(\beta)) = \sum_{i=1}^{n} \left\{ \Delta N.(t_i) \left[\boldsymbol{Z}_i^T(t)\beta - \log\left(\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{Z}_j^T(t)\beta) \right) \right] \right\}$$

Posteriormente se deriva respecto a β para obtener la *función score*, la cual se define de la siguiente manera:

$$U(\beta) = \frac{d}{d\beta}\log(\mathscr{L}(\beta)) = \sum_{i=1}^{n} \Delta N.(t_i)\left[\boldsymbol{Z}_i^T(t) - \frac{\sum_{k=1}^{n} Y_k(t)\exp(\boldsymbol{Z}_k^T(t)\beta)\boldsymbol{Z}_k(t)}{\sum_{k=1}^{n} Y_k(t)\exp(\boldsymbol{Z}_k^T(t)\beta))}\right] \quad (2.1.4)$$

De manera que el estimador de β es solución a la ecuación $U(\beta) = 0$. Por otro lado, la manera en la que se interpretan los coeficientes β es mediante el *cociente de riesgos*, el cuál se define con la siguiente expresión para dos vectores de covariables $\boldsymbol{Z}_1(t)$ y $\boldsymbol{Z}_2(t)$:

$$\frac{h(t|\boldsymbol{Z}_1(t))}{h(t|\boldsymbol{Z}_2(t))} = \exp((\boldsymbol{Z}_1^T(t) - \boldsymbol{Z}_2^T(t))\beta)$$

Representa la comparación del riesgo relativo que existe entre dos individuos descrito a través de sus respectivos valores. De tal modo que se obtiene la interpretación al comparar la función de riesgos del modelo con la función de riesgo de base de la siguiente manera:

$$\log\left(\frac{h(t|\boldsymbol{Z}_1(t))}{h(t|\boldsymbol{Z}_2(t))}\right) = \boldsymbol{Z}_1^T(t)\beta - \boldsymbol{Z}_2^T(t)\beta$$

Así, para un individuo i, si se considera un incremento en alguna de las variables (la primera, sin pérdida de generalidad) manteniendo las demás fijas, se obtendrá que

$$\log\left(\frac{h(t|\boldsymbol{Z}_{i1}+1)}{h(t|\boldsymbol{Z}_{i1})}\right) = \beta_1$$

Si bien, la interpretación parece siempre poder caracterizarse con los coeficientes del vector β, el hecho de que el efecto en el riesgo que describan pueda no ser constante, limita el modelo de riesgos proporcionales paramétrico.

Modelo de riesgos proporcionales no paramétrico

En ocasiones, se observa que el efecto que tienen algunas variables en el modelo de la función de riesgo no es constate a lo largo del tiempo, lo cuál ocasiona que un modelo paramétrico no sea el más adecuado para describir el modelo. Por ello se plantea un vector de coeficientes $\beta(t)$ dependiente del tiempo, de tal modo que la función de riesgo bajo este modelo se representa de la siguiente manera:

$$\lambda(t) = Y(t)\lambda_0(t)\exp(\boldsymbol{X}^T(t)\beta(t)) \quad (2.1.5)$$

El cuál es un modelo no paramétrico, pues un vector β no determina el modelo $\forall\ t$. Cabe mencionar que en la expresión **2.1.5**, la función vector de riesgo de base $\lambda(t)$ es contenida en el factor de proporcionalidad como $\lambda_0(t) = \exp(\beta_0(t))$ [2] (aunque en algunas ocasiones sí se exprese). Obsérvese que el coeficiente $\beta_0(t)$ depende del tiempo, por lo que no se pierde el aporte de la función de riesgo de base dentro del modelo.

[2] Esta es la forma en la que se hace la estimación en la función *timecox()* del paquete *timereg*

La forma en la que se estiman los coeficientes, es a través de la siguiente verosimilitud basada en la expresión **2.1.3**. Al considerar el tiempo $T \in [0, \tau]$ como una variable continua, se expresa la verosimilitud como:

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0}^{\tau} \left(\frac{\exp(\boldsymbol{X}_i^T(t)\beta(t))}{\sum_{j=1}^{n} Y_j(t) \exp(\boldsymbol{X}_j^T(t)\beta(t))} \right)^{\Delta N_\cdot(t)}$$

Siendo continua la variable de tiempo T, se obtiene la siguiente función de log-verosimilitud:

$$\log(\mathscr{L}) \approx \sum_{i=1}^{n} \left\{ \int_0^\tau \boldsymbol{X}_i^T(t)\beta(t) dN_\cdot(t) - \int_0^\tau Y_j(t) \exp(\boldsymbol{X}_j^T(t)\beta(t)) dt \right\}$$

Derivando respecto a $\beta(t)$ se obtiene la función score:

$$\frac{d}{d\beta} \log(\mathscr{L}(\beta)) = U(\beta(t)) = \sum_{i=1}^{n} \left\{ \int_0^\tau \boldsymbol{X}_i^T(t) dN_i(t) - \boldsymbol{X}_i^T(t)\lambda(t)dt \right\} \tag{2.1.6}$$

Y al igualar a 0 la función score anterior, se obtiene el siguiente sistema de ecuaciones:

$$\boldsymbol{X}^T(t)(dN(t) - \lambda(t)dt) = 0$$

El cual no tiene solución analítica dado que $dN(t)$ es un proceso de saltos y $\lambda(t)dt$ es continua. Sin embargo, la estimación se puede realizar mediante la función *timecox()* del paquete *timereg*.

Modelo de riesgos proporcionales semi-paramétrico

Cuando algunas variables tienen un efecto constante a lo largo del tiempo y algunas otras no, se considera el modelo semi-paramétrico, el cuál permite incluir simultáneamente las variables que contribuyen al modelo de manera paramétrica y no paramétrica. De este modo, se define el modelo como:

$$\lambda(t) = Y(t) \exp(\boldsymbol{X}^T(t)\beta(t) + \boldsymbol{Z}(t)\gamma) \tag{2.1.7}$$

Este modelo se compone de dos partes: la parte no-paramétrica que corresponde a las covariables especificadas por la matriz $\boldsymbol{X}(t)$ y el vector $\beta(t) = (\beta_1(t), ..., \beta_p(t))$ en el primer sumando dentro de la función exponencial, y la parte paramétrica que se especifica con $\boldsymbol{Z}(t)$ y $\gamma = (\gamma_1, ..., \gamma_q)$ en el segundo sumando. La estimación de los coeficientes se realiza de un modo similar que en el caso puramente no paramétrico. Se tiene la función score de la expresión **2.1.4** para un valor γ fijo y con ella se inicia el proceso iterativo. Este modelo se considera como una generalización de los dos anteriores, dado que para $p = 0$ se tiene el modelo de riesgos proporcionales simple y para $q = 0$ se tiene el modelo no paramétrico.

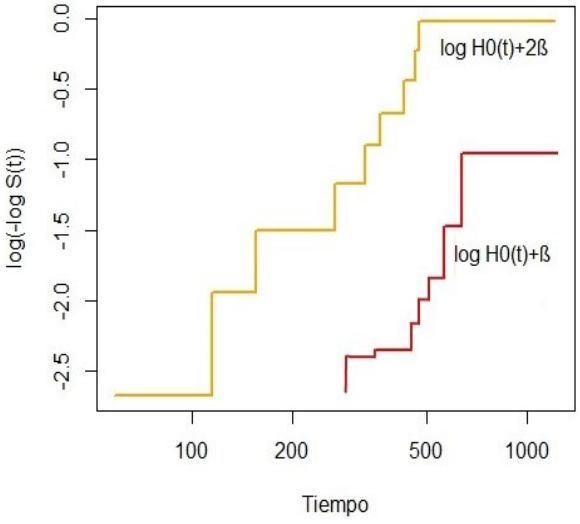
Pruebas de proporcionalidad y evaluación del ajuste

El único supuesto que se hace en el modelo **2.1.7** es que el factor $\alpha(\boldsymbol{Z}^T(t)) = \exp(\boldsymbol{Z}^T(t)\beta(t))$ es proporcional a $\lambda_0(t)$. Dado que las covariables de la matriz $\boldsymbol{Z}^T(t)$ pueden ocasionar desviaciones de este spuesto, la prueba de hipótesis más relevante consiste en verificar si se cumple el supuesto de proporcionalidad del modelo respecto a todas sus covariables. Es decir, se requiere contrastar la hipótesis de que para cada variable es posible ajustar un modelo proporcional, i.e:

$$H_0 : \lambda(t) = Y(t)\lambda_0(t)exp(\boldsymbol{X}_j^T(t)\beta) \quad \text{vs.} \quad H_1 : \lambda(t) \neq Y(t)\lambda_0(t)exp(\boldsymbol{X}_j^T(t)\beta) \tag{2.1.8}$$

Existen varias formas de comprobar la proporcionalidad de un modelo. Alguna de ellas son gráficas y otras más se realizan con pruebas formales de hipótesis. En el caso del modelo paramétrico, la manera más intuitiva es estratificar la estimación de la función de supervivencia $S(t)$ para las distintas variables discretas y continuas. Para las variables continuas, los estratos se definen al clasificar los datos según los cuartiles, para las discretas únicamente sus valores. La prueba se realiza al observar las gráficas de la transformación $\log(-\log(\hat{S}(t)))$ (siendo $\hat{S}(t)$ la estimación de Kaplan-Meier) para cada estratificación, para así juzgar si son paralelas o no. Se espera que las curvas sean paralelas para las variables proporcionales. Esto se puede ver dada la siguiente equivalencia (suponiendo el modelo de riesgos proporcionales):

Figura 2.1: Curvas del estimador de Kaplan-Meier estratificado para los dos niveles de una variable proporcional en algún modelo.



$$\log(-\log(S(t))) = \log(H(t)) = \log(H_0(t)exp(\boldsymbol{X}^T(t)\beta)) = \log H_0(t) + \boldsymbol{X}^T(t)\beta$$

Es decir, si $\boldsymbol{X}(t)$ es una variable discreta, para cada valor representará una constante sobre la curva $\log H_0(t)$, siempre y cuando se cumpla el supuesto del modelo de riesgos proporcionales. En la **fig. 2.1** se ejemplifica la interpretación de proporcionalidad de una variable categórica de dos niveles, cuyas curvas son casi paralelas.

Prueba para el modelo de riesgos proporcionales semi-paramétrico

Sin importar si la variable, cuya proporcionalidad se deseé probar, se asocie a un coeficiente β o $\beta(t)$, una prueba formal para contrastar la hipótesis de proporcionalidad del modelo semi-paramétrico sería expresar los coeficientes de regresión como:

$$\beta_j(t) = \beta_j + \theta_j g_j(t)$$

donde $g_j(t) = log(t)$ (en general $g_j(t)$ puede tomar distintas formas) servirá para probar la hipótesis de proporcionalidad, i.e.

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0 \text{ con } \theta = (\theta_1, ..., \theta_p) \tag{2.1.9}$$

ya que si $\theta_j \neq 0$ para alguna $j = 1, ..., p$, se tendrá una desviación de la hipótesis de proporcionalidad debido a la j-ésima variable:

$$\lambda(t) = \lambda(t) \exp(\beta_j(t)\boldsymbol{X}_j(t))t^{\theta_j} \tag{2.1.10}$$

i.e. un modelo no proporcional, debido al término generado por la variable $\boldsymbol{X}_j(t)$. Así que la prueba de hipótesis de proporcionalidad se puede ver como:

$$T = U^T I^{-1} U = (U_1^T, U_2^T) \begin{pmatrix} I_{11}^{-1} & I_{12}^{-1} \\ I_{21}^{-1} & I_{22}^{-1} \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$$

donde la matriz I de información de Fisher observada se calcula suponiendo la expresión **2.1.10** del modelo de riesgos proporcionales de la siguiente manera:

$$\hat{I}(\beta, \theta) = - \left[\begin{matrix} \frac{\partial^2 \log(\mathscr{L})}{d\beta^2} & \frac{\partial^2 \log(\mathscr{L})}{\partial\beta\partial\theta} \\ \frac{\partial^2 \log(\mathscr{L})}{\partial\beta\partial\theta} & \frac{\partial^2 \log(\mathscr{L})}{\partial\theta^2} \end{matrix} \right] \Bigg|_{(\beta,\theta)=(\hat{\beta},\hat{\theta})} \tag{2.1.11}$$

Por otro lado, el vector score $U = (U_1^T, U_2^T)^T$ tiene dos componentes: la derivada respecto a β y θ. Evaluándolo en la hipótesis nula **2.1.9** de proporcionalidad, $T(\hat{\beta}, 0)$, se obtiene

$$T(\hat{\beta}, 0) = U_2^T(\hat{\beta}, 0)^T I^{22}(\hat{\beta}, 0) U_2(\hat{\beta}, 0) \tag{2.1.12}$$

la cual se distribuye asintóticamente como $\chi^2_{(p)}$. La función *cox.zph()* del paquete *survival* ejecuta esta prueba para cada variable suponiendo proporcionalidad para todas las demás, i.e. desde el punto de vista de cada variable. También incluye el resultado desde el punto de vista global del modelo. Cabe aclarar que esta forma de probar la hipótesis puede ocasionar errores de interpretación. Esto no sólo debido al hecho de que desde el punto de vista de cada variable, suponer simultáneamente que las demás son proporcionales puede llevar a errores, y por si fuera poco, también hay que recordar que en general, el resultado depende de la función que se ocupe para $g_j(t)$, lo cual puede llevar a diversas interpretaciones y resultados. Por ello se tienen otras formas que podrían reparar este problema de interpretación.

Proceso de scores (*Score process*)

Para complementar visualmente la prueba anterior y en el caso de los modelos paramétricos, se tiene otra prueba desarrollada en el análisis de procesos de conteo. El llamado *Score process*, descrito por *Martinussen & Scheike* (2006, pág. 197) es un análisis que se basa en los residuales del tipo score. Al evaluar el estimador $\hat{\beta}$ en la función score de la expresión **2.1.4**, se obtiene

$$U(\hat{\beta}, t) = \sum_{i=1}^{n} \int_0^t \boldsymbol{X}_i(s) d\hat{M}_i(s) \tag{2.1.13}$$

Este último proceso, detallado por *Martinussen & Scheike* (2006, pág 198) es el que se grafica como curvas correspondientes a las hipótesis nula de proporcionalidad. Luego, con el objetivo de evaluar el comportamiento del modelo, se contrastan las curvas simuladas por variables gaussianas G_i con la obtenida con los datos de la muestra sustituidos en la siguiente expresión.

$$n^{-1/2} \sum_{i=1}^{n} \int_0^t (\boldsymbol{X}_i(s) - \frac{\sum_i Y_i(s) exp(\boldsymbol{X}_i^T(s)\hat{\beta}) \boldsymbol{X}_i(s)}{\sum_i Y_i(s) exp(\boldsymbol{X}_i^T(s)\hat{\beta})}) d\hat{M}_i(s) G_i \tag{2.1.14}$$

Tiene sentido comparar estar curvas por que bajo la hipótesis **2.1.8**, las expresiones **2.1.13** y **2.1.14** tienen la misma distribución asintótica. La función de *cox.aalen()* realiza por defecto 50 simulaciones para cada covariable en el modelo[3]. Si la curva obtenida de las observaciones se aleja significativamente del resto de las curvas correspondientes a las simulaciones, se tendría evidencia para rechazar la proporcionalidad de la j-ésima variable. Más aún, la función en R *plot.cox.aalen()* del paquete *timereg* muestra los resultados de la siguiente estadística

$$\sup_{t \in [0,\tau]} |U_j(\hat{\beta}, t)| \tag{2.1.15}$$

basada en el vector $U(\hat{\beta}, t) = (U_1, ..., U_p)$, relativa a cada una de las covariables $j = 1, ..., p$ del modelo ajustado. Al estimar su distribución, se obtiene un p-valor para contrastar la siguiente hipótesis:

$$H_0 : U_j(\hat{\beta}, t) = 0 \text{ vs. } H_1 : U_j(\hat{\beta}, t) \neq 0$$

Esta es otra forma de apoyar gráficamente el análisis de proporcionalidad de las variables y solucionar algunos problemas de las otras pruebas. Sin embargo, no es la única prueba que se debe aplicar, pues incluye el defecto de poder concluir proporcionalidad aún cuando es evidente que en algunos subintervalos de $[0, \tau]$ no se tiene proporcionalidad.

[3]Esta cantidad sólo está limitada por el rendimiento computacional promedio y puede modificarse. Sólo se recomienda este número para observar el contraste con la curva de las observaciones

Alternativas a la proporcionalidad

Cuando no se puede considerar que alguna variable es proporcional, debido a que se rechaza la hipótesis en las pruebas anteriores, lo más adecuado es emplear las siguientes alternativas:

- Estratificar: Cuando se tienen razones para incluir una variable en el modelo, aún cuando esta no es proporcional, una alternativa es estratificar la población con esta variable, de tal modo que se tengan funciones basales distintas.
- Truncar el tiempo: En las pruebas gráficas anteriores es posible observar si existen algún subintervalo de $[0, \tau]$ en donde se cumpla la proporcionalidad de manera significativa. Si esto es verdad, se puede truncar el tiempo de estudio y sólo restringirlo en el subintervalo de tiempo.
- Usar otros modelos: se puede analizar el mismo conjunto de variable con otros modelos, como el modelo de Aalen o el aditivo-multiplicativo, descritos en las **secciones 2.2** y **2.3** .

Estratificación del modelo.

Si se desea partir la población analizada acorde a ciertos estratos, es posible incluir esta información en el modelo. De acuerdo con *Klein & Moeschberger* (2003, pág. 308), este planteamiento se puede expresar como:

$$\lambda_i(t) = \left(\sum_{k=1}^{u} \lambda_{0k}(t) \mathbb{1}_{ik} \right) exp(\boldsymbol{X}_i^T(t)\beta(t) + \boldsymbol{Z}_i^T(t)\gamma)$$

para el k-ésimo estrato con $k = 1, ..., u$ y el i-ésimo individuo con $i = 1, ..., n$.

$$\mathbb{1}_{ik} = \begin{cases} 1 & \text{si el individuo } i \text{ pertenece al estrato } k \\ 0 & \text{en otro caso} \end{cases}$$

En el modelo paramétrico, se tiene el mismo vector de coeficientes $\beta \ \forall t$, pero la función de riesgo de base es diferente. En el modelo no paramétrico, dado que se considera por simplicidad que $\lambda_0(t) = exp(\beta_0(t))$, se tendrán u estimaciones $\beta_{0j}(t)$ para cada estrato $k = 1, ..., u$. La estratificación se puede realizar para cualquier tipo de variable, sin importar que sea discreta o continua. En el caso de las continuas, se realiza de acuerdo a la elección de cuartiles, mientras que las variables discretas se estratifican según sus niveles.

Estimación de la función de supervivencia.

Conforme a la expresión **2.1.7** del modelo semi-paramétrico, se puede obtener la función de supervivencia expresada como:

$$\hat{S}_0(t) = exp(-\hat{\Lambda}(t)) = exp(-\int_0^t exp(\boldsymbol{X}_0^T\hat{\beta}(s) + \boldsymbol{Z}_0^T\hat{\gamma})ds) \tag{2.1.16}$$

para un grupo de variables con determinados valores $\boldsymbol{X}_0$ y $\boldsymbol{Z}_0$. El problema con el modelo es que no es sencillo obtener una gráfica de la función de supervivencia dado que la integral en el argumento de la función exponencial no es una expresión que se pueda calcular fácilmente. Para estos casos, se puede optar por mostrar la gráfica de la función de riesgo como alternativa visual del modelo.

2.2. Modelo de riesgos aditivos

Estos modelos son descritos por Odd Aalen y consisten en modelos en los cuales los coeficientes de las covariables pueden tener un efecto variable sobre la función de riesgo λ o pueden ser constante a través del tiempo. El modelo aditivo es otra forma de modelar la función de riesgo de los individuos. Para este modelo, no se considera una función de riesgo de base, sino una regresión en la que los coeficientes dependen del tiempo. Así entonces, de acuerdo con *Aalen* (2008, pág. 154), se tiene:

$$\lambda(t) = Y(t)\boldsymbol{X}^T(t)\beta(t) \tag{2.2.1}$$

Parte importante del análisis es el efecto que tienen las variables a lo largo del intervalo de tiempo de análisis $[0, \tau]$. Para poder observar esto se define la *función vector de coeficientes de regresión acumulada* $(B_1(t), ..., B_n(t))$ como:

$$B(t) = \int_0^t \beta(s)ds$$

De este modo, la estimación de estos coeficientes se obtendrá de la *descomposición de Doob-Meyer* **1.0.4** sustituyendo en la expresión **2.2.1**. Así el estimador para cada tiempo t es:

$$\hat{B}(t) = \int_0^t \boldsymbol{X}^{-1}(s)dN(s)$$

La estimación y observación de $B(t)$ es importante tanto en el modelo de riesgos aditivos como el de riesgos proporcionales, ya que una gráfica de $B(t)$ vs. t permitiría observar el comportamiento de los coeficientes $\beta(t)$ de cada covariable para poder evaluar qué tanto varían a lo largo del tiempo.

Modelo semi-paramétrico

Análogamente al modelo de riesgos proporcionales, se pueden tener también algunas covariables cuyo efecto varía a través del tiempo, mientras otras no. Por ello, *Martinussen & Scheike* (2008, pág. 135) definen el modelo de riesgos semi-paramétrico como:

$$\lambda(t) = Y(t)(\boldsymbol{X}^T(t)\beta(t) + \boldsymbol{Z}^T(t)\gamma)$$

donde se separa la matriz de observaciones en $\boldsymbol{X}^T(t)$, una matriz $p \times n$-dimensional cuyas covariables tienen un efecto variable $\beta(t)$ p-dimensional y $\boldsymbol{Z}^T(t)$ cuyas covariables tienen un efecto constante γ q-dimensional. Esta forma es conveniente para poder contrastar las variables invariantes de aquellas que varían con el tiempo en el modelo.

Los métodos para la inferencia para este modelo, descritos por *Martinussen & Scheike* (2006, pág. 127) sobre los datos es similar al modelo no paramétrico anterior. Los estimadores para los coeficientes γ se calculan de la siguiente manera:

$$\hat{\gamma} = \left(\int_0^\tau \boldsymbol{Z}^T(t)\boldsymbol{H}(t)\boldsymbol{Z}(t)dt \right)^{-1} \int_0^\tau \boldsymbol{Z}^T(t)\boldsymbol{H}(t)dN(t)$$

donde $\boldsymbol{H}(t) = I - \boldsymbol{X}(t)\boldsymbol{X}^{-1}(t)$. Y para el vector de coeficientes acumulados $B(t)$:

$$\hat{B}(t) = \int_0^\tau \boldsymbol{X}^{-1}(s)(dN(s) - \boldsymbol{Z}(s)\hat{\gamma}ds)$$

La forma en la que se realizan las pruebas de hipótesis para el modelo aditivo se enfocarán en analizar la variabilidad del efecto de las covariables en el riesgo modelado, i.e la prueba de Kologorov-Smirnov contrasta las siguientes hipótesis:

$$H_0 : B(t) = \gamma t \ \text{ vs. } \ H_1 : B(t) \neq \gamma t$$

En la **sección 2.5** se describe el funcionamiento de esta prueba. La forma más apropiada de realizar el análisis, es suponer el modelo no paramétrico y mediante las pruebas de hipótesis de efecto constante, descartar las variables con tales efectos, tal y como se ha hecho hasta ahora.

Estratificación del modelo aditivo

Análogo al modelo de riesgos proporcionales, para el modelo aditivo se tiene también una forma de estratificar la población. Para ello se sigue la misma idea de plantear una función de riesgo de base para cada estrato. Siguiendo esta idea, se puede expresar la función de riesgo como:

$$\lambda_i(t) = \sum_{k=1}^{u} \lambda_{0i}(t)\mathbb{1}_{ik} + \boldsymbol{X}_i^T(t)\beta(t) + \boldsymbol{Z}_i^T(t)\gamma$$

Se obtiene así una función de riesgo de base para cada estrato k, la cual se representa como

$$\lambda_0(t) = \sum_{i=1}^{u} \lambda_{0i}(t)\mathbb{1}_{ik} + \beta_0(t)$$

Estimación de la función de Supervivencia

Para la estimación de la función de supervivencia, se retoma la definición de la función de riesgo, suponiendo que las variables X_0 y $\boldsymbol{Z}_0$ no dependen del tiempo:

$$\lambda_0(t) = \boldsymbol{X}_0^T\beta(t) + \boldsymbol{Z}_0^T\gamma$$

de esta función de riesgo, se puede escribir la función de supervivencia de la siguiente forma:

$$S_0(t) = S_0(B, \gamma, t) = exp(-\boldsymbol{X}_0^T B(t) - \boldsymbol{Z}_0^T t\gamma)$$

que claramente se puede estimar como:

$$\hat{S}_0(t) = S_0(\hat{B}, \hat{\gamma}, t) = exp(-\boldsymbol{X}_0^T \hat{B}(t) - \boldsymbol{Z}_0^T t\hat{\gamma})$$

Con este modelo no se tienen los problemas del modelo multiplicativo, ya que se tiene la expresión anterior analíticamente.

2.3. Modelo de riesgos aditivo-multiplicativo

Las variables pueden conformar un modelo adecuado que se ajuste a las observaciones mediante el modelo de riesgos proporcionales o en el modelo aditivo. Para complementar el análisis con los modelos anteriores, es posible incorporar el efecto de algunas variables proporcionales con algunas otras aditivas en el mismo modelo. De esta idea, se desprende el modelo Aditivo-Multiplicativo, el cuál ha sido estudiado por diversos autores. A decir verdad, no hay un único modo de definir este modelo, pues existen diversas formas de representarlo. Estos modelos pueden representar diversas características e interpretaciones según la forma en la que se presenten.

Un modelo general, propuesto por *Dabrowska* (1997), considera una función basal $\alpha(t, \boldsymbol{X})$ que es una función dependiente del tiempo para un sujeto con las características basales de $\boldsymbol{X}$. La forma de modelar esta función de intensidad se representa como:

$$\lambda(t) = Y(t)\{\alpha(t, \boldsymbol{X})\}exp(\boldsymbol{Z}^T(t)\beta) \tag{2.3.1}$$

La idea detrás de esta forma de modelar la función de riesgo es un modelo similar al de riesgos proporcionales, en el que se considera que la función de riesgo de base varía a lo largo del tiempo de acuerdo a un conjunto de covariables $\boldsymbol{X}$, las cuales pueden representar estratos en el caso de que sean categóricas.
Para representar la función $\alpha(t, \boldsymbol{X})$ también hay distintas maneras, aunque en el presente trabajo se considerará que la función de riesgo será aquella descrita por *Scheike & Zhang* (2002), la cual se conoce también como modelo de riesgos aditivo-multiplicativo. Haciendo $\alpha(t, \boldsymbol{X}) = \boldsymbol{X}_i^T(t)\beta(t)$ se define de la siguiente manera:

$$\lambda(t) = Y(t)\{\boldsymbol{X}^T(t)\beta(t)\}exp(\boldsymbol{Z}^T(t)\gamma) \tag{2.3.2}$$

donde $Y(t)$ es el vector de dimensión n que indica la exposición al riesgo, $\beta(t)$ es el vector de coeficientes dependientes del tiempo de dimensión q, mientras que γ es el vector de coeficientes constantes. Esta forma brinda una representación aditiva para las variables de $\boldsymbol{X}(t)$ que tienen un efecto variable en la función de riesgos y una representación multiplicativa para las variables de $\boldsymbol{Z}(t)$ que tienen un efecto constante, de dimensiones q y p respectivamente.
El modelo **2.3.2** es un caso particular del modelo de Dabrowska. Pero para ver el modelo también como un caso general del modelo de riesgos proporcionales estratificado, se expresa la función de riesgo de base estratificada como $\alpha(t, \boldsymbol{X}) = \boldsymbol{X}^T(t)\beta_0(t)$, lo cuál equivale a la función de riesgo de base para el modelo de riesgos proporcionales estratificado si $\boldsymbol{X}(t)$ es una variable discreta con u niveles o estratos.

Lo interesante de estos modelos, es que brindan completa libertad para las covariables a ser agregadas al modelo, pues algunas pueden incluirse aditiva o multiplicativamente según las hipótesis que se cumplan. Así que una vez que se hayan realizado las pruebas de hipótesis de proporcionalidad o efecto constante, este modelo brinda flexibilidad para discriminar qué variables ingresan al modelo de forma aditiva o multiplicativa según las suposiciones que cumplan. En R, la función *cox.aalen()* del paquete *timereg* permite etiquetar cada covariable con el atributo de ser aditiva o multiplicativa. La forma en la que se hacen las estimaciones para los coeficientes de estos modelos se incluye en *Martinussen & Scheike* (2005, pág. 252).

2.4. Significancia y efecto constante de las covariables del modelo

Pruebas de significancia

Modelo paramétrico

Lo que se desea probar es la existencia de una relación entre las variables y la función de intensidad del modelo, a eso se refiere la significancia de las variables. Un supuesto en estas pruebas es que las coeficientes $\beta_j(t) = \beta_j \; \forall \, t, j = 1, .., p$. Así que con el fin de contrastar la hipótesis de significancia para la j-ésima entrada del vector de coeficientes β del modelo, se plantea la siguiente prueba de hipótesis:

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \;\; j = 1, \ldots, p$$

para la cual se tiene la siguiente estadística basada en la j-ésima entrada de la diagonal de la matriz de información de Fisher observada definida en la expresión **2.1.11**:

$$Z = \frac{\hat{\beta}_j}{\sqrt{I(\hat{\beta})^{-1}_{jj}}} \sim N(0, 1)$$

Se rechaza la hipótesis nula si $Z > |z_{1-\alpha/2}|$. Para modelos anidados, si se desea contrastar la hipótesis de significancia simultáneamente de r covariables en un modelo con covariables $\beta_1, ..., \beta_p, \beta_{p+1}, ..., \beta_{p+r}$, i.e.

$$H_0 : \beta_{p+1} =, ..., = \beta_{p+r} = 0 \text{ vs } H_1 : \beta_j \neq 0 \text{ para alguna } j = p + 1, ..., p + r$$

con respecto a un modelo anidado con $(\beta_1, ..., \beta_p)$, se tiene el siguiente cociente de verosimilitudes, el cuál tiene una distribución asintótica $\chi^2_{(r)}$.

$$Z = -2 \log \left(\frac{L(\beta_1, ..., \beta_p)}{L(\beta_1, ..., \beta_p, \beta_{p+1}, ..., \beta_{p+r})} \right)$$

Análogamente, si al comparar resulta $Z > \chi^2_{(r),1-\alpha}$, se rechaza con $100\alpha\,\%$ de significancia la hipótesis nula. Obsérvese que con esta prueba también se puede comprobar la hipótesis de significancia del modelo entero al comparar el modelo completo con aquél donde todos los coeficientes del vector $(\beta_1, ..., \beta_p)$ son nulos. Sin embargo, hay que recordar que cuando no se tiene evidencia para hacer el supuesto de que $\beta(t)$ es constante a lo largo del tiempo, lo más adecuado es recurrir a los métodos para modelos no paramétricos descritos a continuación.

Pruebas para modelos no paramétricos

La siguiente prueba se basa en la prueba de hipótesis anterior de significancia pero para coeficientes que varían a lo largo del tiempo, i.e. se desea contrastar

$$H_0 : \beta_j(t) = 0 \text{ vs } H_0 : \beta_j(t) \neq 0 \tag{2.4.1}$$

para la j-ésima entrada del vector de coeficientes $\beta(t)$, pero considerando que los coeficientes $\beta(t)$ dependen del tiempo, lo más adecuado es utilizar la siguiente estadística:

$$\sup_{t\in[0,\tau]} \left| \frac{\hat{B}_j(t)}{\hat{\mathbf{\Phi}}_{jj}(t)} \right| \tag{2.4.2}$$

donde $\hat{\mathbf{\Phi}}$ es una estimación de la varianza del estimador $\hat{B(t)}$, la cual se expresa como:

$$\hat{\mathbf{\Phi}} = n \int_0^\tau \boldsymbol{X}^{-1}(s)\text{diag}(dN(s))(\boldsymbol{X}^{-1}(s))^T \tag{2.4.3}$$

según *Martinussen & Scheiek* (2006, pág. 112). La función *aalen()* del paquete *timereg* realiza la estimación correspondiente a **2.4.3** y para contrastar la hipótesis nula **2.4.1**, estima el p-valor de la distribución de **2.4.2** para determinar si es significativo el coeficiente asociado a la j-ésima variable. Dado que con los intervalos de confianza también se puede inferir la significancia de los parámetros, con la función de covarianzas de la expresión **2.4.3** pueden construirse intervalos junto con las curva de la estimación de los coeficientes $B(t)$.

Efecto constante

Prueba de Kolmogorov-Smirnov

La siguiente prueba es un caso particular de la prueba de Kolmogorov-Smirnov. En este contexto, es de interés probar la siguiente hipótesis para una constante γ:

$$H_0 : B_j(t) = \gamma t \quad \text{vs.} \quad H_0 : B_j(t) \neq \gamma t \tag{2.4.4}$$

Y según *Martinussen & Scheike* (2006, pág. 116), para un tiempo de observación τ y dada la estadística de la siguiente expresión para la j-ésima entrada del vector $B(t)$:

$$n^{1/2} \sup_{t\in[0,\tau]} \left| \hat{B}_j(t) - \hat{B}_j(\tau)\frac{t}{\tau} \right| \tag{2.4.5}$$

se pretende aproximar su distribución a la *distribución de Kolmogorov*. La idea detrás de esto es que $\hat{B}_j(\tau)/\tau$ funciona como una estimación de la constante γ bajo la hipótesis nula **2.4.4**. Se compara entonces, el valor de la expresión **2.4.5** con la aproximación de la distribución de Kolmogorov, que bajo la hipótesis nula **2.4.4** es estimada por:

$$\Delta_1(t) - \Delta_1(\tau)\frac{t}{\tau} \tag{2.4.6}$$

donde, bajo condiciones generales, y dadas $\{G_1, ..., G_n\} \sim N_n(0, 1)$ independientes y distribuidas normal estándar, se tiene que $n^{1/2}(\hat{B}_j(t) - B_j(t))$ tiene la misma distribución límite que

$$\Delta_1(t) = n^{-1/2} \sum_{i=1}^{n} \hat{\epsilon}_i(t) G_i \tag{2.4.7}$$

condicionada a los datos $(N_i, Y_i, \boldsymbol{X}_i)$ para los individuos $i = 1, ..., n$. Donde las estimaciones $\hat{\epsilon}_i(t)$ se realizan mediante:

$$\hat{\epsilon}_i(t) = \int_0^\tau (n^{-1} \boldsymbol{X}^T(s) \boldsymbol{X}(s))^{-1} \boldsymbol{X}_i(s) d\hat{M}_i(s)$$

De este modo, bajo la hipótesis nula **2.4.4**, la distribución de Kolmogorov se aproxima mediante la generación de simulaciones de la expresión **2.4.7**. Así, se compara entonces el valor obtenido en **2.4.5** con la distribución de Kolmogorov estimada y se rechaza conforme a cierto nivel de significancia dado.
La justificación de este razonamiento y más acerca de la distribución de Kolmogorov pueden hallarse en *Marsaglia & Tsang* (2003).

Criterio de Cramér-von Mises

También existe otro criterio para hacer la prueba de la misma hipótesis de **2.4.4**. Se tiene entonces, el criterio de Cramér-von Mises. Funciona de un modo parecido y se encuentra igualmente presentado por las funciones del paquete *timereg* para el objetivo de comparar ambas pruebas y elegir las covariables más significativas. Para esta prueba se remplaza **2.4.5** por la siguiente estadística:

$$n \int_0^\tau \left(\hat{B}_j(t) - \hat{B}_j(\tau) \frac{t}{\tau} \right)^2 dt$$

El único inconveniente de estas pruebas es que consideran información de todo el intervalo $[0, \tau]$ y concluyen sobre la significancia o efecto dependiente del tiempo para todo el intervalo aún cuando puede ser evidente que no sucede lo mismo para un subintervalo $[0, \tau_1]$. Por ello es importante el apoyo gráfico que brinda la misma paquetería *timereg* en R.

Proceso de prueba (Test Process)

Basándose en la prueba de Kolmogorov-Smirnov, se tiene la siguiente prueba gráfica para contrastar la hipótesis nula **2.4.4** de efecto constante para la j-ésima covariable del modelo, la cual se describe por *Martinussen & Scheike* (2006, pág. 122). El objetivo es observar el proceso definido por:

$$\hat{B}_j(t) - \hat{B}_j(\tau) \frac{t}{\tau} \tag{2.4.8}$$

con los datos de la muestra y compararlo con las realizaciones del proceso definido por la expresión **2.4.6** bajo la hipótesis nula de efecto constante. Conforme a la sección anterior,

para cada realización se simulan n variables aleatorias normal estándar (50 es la cantidad estandarizada) y usando la estimación de la expresión **2.4.7**, se obtienen los valores simulados de **2.4.6** (aproximación de la distribución de Kolmogorov). De tal modo que se comparan las curvas de las simulaciones con la curva obtenida de las observaciones y se observa si existen intervalos de tiempo donde se aparte considerablemente la curva del modelo con las realizaciones. Si esto sucede, se considera que se tiene un efecto dependiente del tiempo por parte de esa variable. El apoyo gráfico se recomienda junto con la prueba de Kolmogorov descrita en la sección anterior.

Grupos de riesgo

Una cuestión interesante es investigar si hay alguna forma de seleccionar a los individuos para formar grupos de riesgo. Los grupos de riesgo son clasificaciones definidas para distinguir aquellos elementos de la muestra que son más propensos a una falla.
Existen varias formas de clasificar los individuos de la muestra, pero algunos autores como *Fleming & Harrington* (1991, pág 190) o *Machin & Parmar* (2006, pág 187) utilizan el *Índice de pronóstico* (*IP*). Este representa una medida de riesgo para cada paciente, dependiendo de los valores de sus características en cada variable. Con los terciles de esta cantidad, se establece cuáles son los individuos con bajo, medio y alto riesgo, dependiendo proporcionalmente del cuartil en que se encuentre. Este índice se propone de diversas maneras de conforme al modelo que represente. Para el modelo multiplicativo, *Machin & Parmar* consideran el índice de pronóstico (IP) como la función de riesgo sin el riesgo de base, ya que debe ser la misma para todos los individuos.

$$IP = exp(\boldsymbol{Z}^T\beta)$$

Pero dado que la función exponencial es monótona, los cuartiles del IP delimitarían los mismos grupos de individuos que los cuartiles de $log(IP)$. Por ello, para el modelo multiplicativo y aditivo se calculará el Índice de pronóstico con la siguiente expresión, de acuerdo con *Fleming & Harrington*.

$$IP = \boldsymbol{Z}^T\beta \tag{2.4.9}$$

Para el modelo aditivo-multiplicativo no se tiene una propuesta de índice de pronóstico que funcione, pues no se puede calcular para modelos no-paramétricos donde los coeficientes $\beta(t)$ varían con respecto al tiempo. Ya que de este modo, el riesgo varía a lo largo del tiempo para el mismo individuo. Por ello, para que se pueda calcular esta cantidad, debe considerarse en el índice sólo coeficientes constantes β.

2.5. Análisis de residuales

Los residuales del modelo pueden proveer información acerca de la bondad de ajuste del modelo. Ya sea para evaluar la proporcionalidad de las covariables en el modelo de riesgos proporcionales o alguna transformación de ellas, identificar puntos de influencia o detectar outliers en el modelo. Para su aplicación en los modelos de riesgos proporcionales paramétricos, se presentan los residuales de martingalas, score y devianza haciendo uso de la paquetería *survival*, cada uno con distintas características que se aprovechan en diversas situaciones. Por otro lado, para los modelos de riesgo no paramétricos (proporcionales, aditivos, aditivos-multiplicativos), se muestran los residuales acumulativos (*cumulative residuals*) haciendo uso del paquete *timereg*.

Residuales de Martingalas

La martingala $N(t)$ en la *descomposición de Doob-Meyer* **1.0.4** puede interpretarse como una especie de ruido en un modelo de regresión estándar (*Variable respuesta* = *Señal* + *Ruido*). De tal modo que se obtiene que:

$$M(t) = N(t) - \Lambda(t) = N(t) - \int_0^t \lambda(s)ds \tag{2.5.1}$$

donde $\Lambda(t)$ es el vector compensador cuya forma depende del modelo a analizar, sea multiplicativo, aditivo o aditivo-multiplicativo. Dado que $\forall\ i\ M_i(t) = N_i(t) - \Lambda_i(t)$, se puede concluir que el rango de valores de los residuales son, en este caso, $(-\infty, 1]$. Esto ya que $N_i(t)$ sólo puede tomar el valor de 1 y $\Lambda_i(t)$ no está restringida.
La forma en la que se interpretan estos residuales es como un exceso o disminución de la mortalidad esperada. Esto porque el número de fallas $N(t)$, visto como proceso de conteo, se rige por su compensador $\Lambda(t)$. De tal modo que expresando la martingala $M(t) = N(t) - \Lambda(t)$, se tendrá una tasa de fallas excesiva si $M(t) > 0$, o una disminución de la misma si $M(t) < 0$. Con esto es posible evaluar la discrepancia del ajuste del modelo con respecto a las observaciones.

Con este fin, para cada una de las observaciones de la muestra se evalua el proceso $M(t)$ en el tiempo τ en el que termina el intervalo de observación, y si t_i es el tiempo de falla del i-ésimo individuo, entonces $M(\tau)$ queda expresado para cada observación i como:

$$M_i(\tau) = \delta_i - \int_0^{t_i} \lambda(s)ds$$

En donde δ_i corresponde a la función indicadora de falla del individuo i (0 si hubo censura o 1 si hubo falla). Estos son los residuales de martingala que se grafican con el fin de evaluar el modelo. La estimación se realiza remplazando los valores de β con sus respectivas estimaciones, y suponiendo el caso paramétrico:

$$\hat{M}_i(\tau) = \delta_i - \int_0^{T_i} Y_i(s)exp(\boldsymbol{Z}_i^T(s)\hat{\beta})d\hat{\Lambda}_0(s)$$

Estos residuales se asemejan a los que se usan en el análisis de regresión, pues describen la discrepancia entre los datos y el modelo. Y a pesar de no ser simétricos, en la siguiente sección relativa a los residuales de devianza se muestra cómo corregir este inconveniente.
Así pues, para evaluar el ajuste del modelo, se puede observar la gráfica de los residuales de martingala vs. el índice de pronóstico para cada tiempo t con el fin de evaluar la forma de la función de enlace del modelo. Ejemplos de esto se muestran en la evaluación de las transformaciones de las variables del modelo de riesgos proporcionales paramétrico del **capítulo 3**.

Como mencionan *Fleming & Harrington* (1991, pág. 165), es posible intuir una transformación adecuada para la j-ésima covariable a través de la gráfica de los residuales de martingala. Los

residuales de esta gráfica corresponderá al modelo que incluye todas las covariables exceptuando la j-ésima vs. los valores de la j-ésima covariable. Para observar esto, se parte de la siguiente expresión de la función de intensidad:

$$\lambda(t) = h(X_j)Y(t)exp(\boldsymbol{Z}^T(t)\beta)\Lambda_0(t) \tag{2.5.2}$$

donde X_j es el j-ésimo vector covariable cuya transformación g se investiga. Puede escribirse g de tal modo que $g(X_j) = exp(f(X_j))$. Calculando la esperanza $E[\lambda(t)|\boldsymbol{Z}(t)]$ de la expresión **2.5.2** se puede escribir:

$$d\tilde{\Lambda}(t) = exp(\boldsymbol{Z}^T\beta)\frac{E\left[g(X_j)Y(t)|\boldsymbol{Z}(t)\right]}{E\left[Y(t)|\boldsymbol{Z}(t)\right]}d\Lambda_0(t) = exp(\boldsymbol{Z}^T(t)\beta)\bar{g}(t,\boldsymbol{Z})d\Lambda_0(t) \tag{2.5.3}$$

Ahora, calculando $E[\hat{M}(t)|X_j]$ de la expresión **2.5.1** y sustituyendo **2.5.3**, se obtiene lo siguiente:

$$E[\hat{M}(t)|X_j] = E[\hat{N}(t)|X_j] + E\left[-\int_0^t Y(s)exp(Z^T\hat{\beta}(t))\bar{g}(s,\boldsymbol{Z})d\Lambda_0(s)|X_j\right] +$$
$$E\left[\int_0^t Y(s)\{d\tilde{\Lambda}(s) - exp(\boldsymbol{Z}^T\hat{\beta}(t))d\hat{\Lambda}_0(s)\}|X_j\right] = \text{sumando 1} + \text{sumando 2} + \text{sumando 3} \tag{2.5.4}$$

Fleming & Harrington (1991, pág. 166) demuestran que el tercer sumando de la expresión **2.5.4** tiende a cero casi seguramente si $n \to \infty$ y $Cov(X_j, Z(t)) \to 0$. En cuanto al segundo sumando, para un tiempo t_0 dado y haciendo $\bar{g}(s) = E[\bar{g}(s,\boldsymbol{Z})]$ se puede expresar lo siguiente:

$$\text{Sumando 2} = -E\left[\int_0^t Y(s)exp(\boldsymbol{Z}^T(t)\beta)\frac{\bar{g}(t_0)}{g(X_j)}g(X_j)d\Lambda_0(s)|X_j\right] +$$
$$E\left[\int_0^t Y(s)exp(\boldsymbol{Z}^T(t)\beta)[\bar{g}(t_0) - \bar{h}(s,\boldsymbol{Z}(t))]d\Lambda_0(s)\right]$$

De lo cual, se obtiene lo siguiente:

$$E[\hat{M}(t)|X] \approx \left(1 - \frac{\bar{g}(t_0)}{g(X_j)}\right)E[\hat{N}(t)|X_j] + R(t,X_j)$$

donde $R(t,X_j)$ es un término residual y $\bar{g}(t_0) = E[g(X_j)Y(t_0)|\boldsymbol{Z}(t)]/E[Y(t_0)|\boldsymbol{Z}(t)]$. Lo interesante es que la expresión anterior se puede interpretar como:

$$E[\text{\# Fallas excedentes }] \approx (\text{1-Función de riesgo})E[\text{\# Eventos }]$$

Manipulando la misma expresión, se concluye que:

$$-log\left(1 - \frac{E[\hat{M}(t)|X_j]}{E[\hat{N}(t)|X_j]}\right) \approx f(X_j) - \bar{f}(t,X_j)$$

siendo $\bar{f}(t,X_j)$ la función que se ajustaría en la gráfica con los datos de la variable X_j. Finalmente, mediante una aproximación de Taylor y haciendo tender $t \to \infty$ se obtiene que:

$$E[\hat{M}(t)|X_j] \approx (f(X_j) - \bar{f}(X_j))c$$

siendo c =(# de eventos)/(# de individuos total). Con ello se concluye que si X_j no está fuertemente correlacionada con $\boldsymbol{Z}$, una curva ajustada de los residuales $\hat{M}$ del modelo omitiendo X_j vs. los valores de X_j aproximarían la transformación f correcta para $\boldsymbol{X}_j$.

Residuales de devianza

El problema con los residuales de las martingalas es que no son simétricos. Esto representa una dificultad al interpretarlos, pues superiormente están acotados mientras que inferiormente no, lo cuál puede generar problemas de interpretación. Por ello se tiene la transformación sobre las martingalas dada por los residuales de devianza, los cuales se derivan, según la literatura de Modelos Lineales Generalizados por *McCullagh & Nelder* (1989, pág. 24), de la devianza definida como el siguiente cociente de verosimilitud generalizado:

$$D = 2\{\log \mathscr{L}(saturado) - \log \mathscr{L}(\hat{\beta})\} \tag{2.5.5}$$

En el cuál, la verosimilitud correspondiente al modelo saturado se refiere a aquel en el cuál el vector β estimado varía por cada tiempo observado, de tal forma que el modelo se ajusta con un estimación de $\hat{\beta}_i$ para cada observación i. Usando la aproximación $log(x) \approx x - 1$ la expresión **2.5.5** se puede expresar como:

$$D = 2 \sup_{\hat{\beta}_i} \sum_{i=1}^{n} \left[\int_0^\infty \boldsymbol{Z}_i^T \hat{\beta}_i dN_i(s) - \int_0^\infty Y_i(s) exp(\boldsymbol{Z}_i^T \hat{\beta}_i) d\Lambda_0(s) \right]$$

$$- \sum_{i=1}^{n} \left[\int_0^\infty \boldsymbol{Z}_i^T \hat{\beta} dN_i(s) - \int_0^\infty Y_i(s) exp(\boldsymbol{Z}_i^T \hat{\beta}) d\Lambda_0(s) \right] \tag{2.5.6}$$

Y dado que se trata de un cociente de verosimilitudes generalizado, el supremo correspondiente a **2.5.6** toma su valor cuando se cumple que $\forall\ i$ individuo:

$$\int_0^\infty Y_i(s) exp(Z_i^T \hat{\beta}_i) d\Lambda_0(s) = \int_0^\infty dN_i(s)$$

Lo cual, según *Fleming & Harrington* (1991, pág. 168) lleva a la siguiente equivalencia, que es la definición con que se calculan los residuales transformados:

$$d_i = -sign(M_i) \sqrt{-2(M_i(t) + \delta_i log(1 - M_i(t)/\delta_i)} \tag{2.5.7}$$

donde δ_i es la función indicadora de la falla del i-ésmo individuo. Con esta transformación, se tienen residuales que toman valores en $(-\infty, \infty)$, análogamente se pueden graficar vs. las covariables o el índice de pronóstico con el fin de evaluar el ajuste de los datos al modelo. La distribución de los residuales se asemeja a la distribución normal.

Residuales de Score

Los residuales de Score están definidos por la ecuaciones del vector Score para estimar los parámetros β en cada modelo. Su origen es a partir de la función de verosimilitud en **2.1.3** usada para estimar los coeficientes del modelo:

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left(\frac{exp(\boldsymbol{Z}_i^T(t)\beta)}{\sum_i Y_i(t) exp(\boldsymbol{Z}_i^T(t)\beta)} \right)^{\Delta N.(t)}$$

Luego de derivar respecto a β y sustituir las estimaciones en los valores correspondientes, como muestran *Martinussen & Scheike* (2006, pág. 182) se obtiene:

$$U(\beta, t) = \sum_{i=0}^{n} \int_0^t \left[\boldsymbol{Z}_i(s) - \frac{\sum_{l=0}^{n} Y_i(s) exp(\boldsymbol{Z}_l^T(s)\hat{\beta}) \boldsymbol{Z}_l(s)}{\sum_{l=0}^{n} Y_i(s) exp(\boldsymbol{Z}_l^T(s)\hat{\beta})} \right] d\hat{M}_i(s) \tag{2.5.8}$$

Estos residuales tienen algunas aplicaciones especialmente para el análisis de la proporcionalidad de un modelo. En el caso del modelo de riesgos proporcionales, la función Score en la ecuación anterior se puede expresar de la siguiente manera, evaluando en el estimador $\hat{\beta}$ y $\hat{M}(t)$ para cada tiempo t:

$$U(\hat{\beta}, t) = \sum_{i=0}^{n} \int_0^t (\boldsymbol{Z}_i(s) - \bar{\boldsymbol{Z}}(s)) d\hat{M}(s) \tag{2.5.9}$$

donde cada para cada contribución i se tiene una integral de Riemann-Stieltjes de las discrepancias de los valores esperados y las observaciones ($\boldsymbol{Z}(t)$ y $\bar{\boldsymbol{Z}}(t)$ respectivamente). Los valores esperados corresponden a las medias ponderadas de $\boldsymbol{Z}_l(s)$ con los pesos dados por la función de intensidad del modelo de riesgos proporcionales, tal y como se muestra en la expresión **2.5.8**. Así, si se observan valores de $\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)$ muy grandes para varios individuos i, se tendrá evidencia para sospechar que el modelo no es proporcional. También pueden identificarse algunos puntos de influencia de la misma manera cuando se observa el mismo efecto para contados individuos.

Residuales de martingalas acumulativos

Para los modelos no-paramétricos o semi-paramétricos, se tiene otro contraste de hipótesis para estudiar el ajuste y poder observar la discrepancia entre las estimaciones $N(t)$ y el modelo $\Lambda(t)$, basado en los residuales de martingala. Este análisis es útil porque permite estudiar de otro modo el comportamiento de los residuales al investigar la bondad del ajuste por cada variable, pudiendo evaluar las transformaciones aplicadas a estas. Estos residuales definen un proceso dado por la expresión

$$\hat{M}_K(t) = \int_0^t \boldsymbol{K}^T(s) d\hat{M}(s) \tag{2.5.10}$$

donde la matriz $\boldsymbol{K}(t) = (K_1(t), ..., K_n(t))^T$ de dimensión $n \times k$ define k estratos o niveles (para las variables categóricas) para analizar cada variable. Es decir, especifica un grupo de

datos a usar para la estimación de los residuales. Generalmente se toman los cuartiles para estratificar y esa información se almacena en la matriz $\boldsymbol{K}(t)$. Si se quiere tener un panorama del comportamiento de cada variable continua, se puede preceder de la matriz $\boldsymbol{K(t)}$. Esas opciones se pueden especificar en la función *cum.residuals()* del paquete *timereg*.

Al graficar estos residuales, se tiene una visualización de las sumas de la expresión **2.5.1** que representa $N(t) - \Lambda(t)$ (*#fallas observadas−#fallas predichas*). Con ello se pueden visualizar en el tiempo t cuantas fallas de más se predijeron ($N(t) < \Lambda(t)$), si la curva se muestra por debajo del 0, o cuantas fallas sucedidas no se predijeron ($N(t) > \Lambda(t)$) si la curva se sitúa por arriba del 0.

Además de las gráficas, puede ser interesante contrastar formalmente las hipótesis

$$H_0 : \hat{M}_K(t) = 0 \quad \text{vs.} \quad H_1 : \hat{M}_K(t) \neq 0 \tag{2.5.11}$$

ya sea de los residuales calculados con todos los datos de o aquellos dados por las p estratificaciones. En dado caso, la hipótesis anterior se contrastaría para cada estrato de residuales $M_{K_p}(t)$. Esta prueba se basa en la estadística

$$\sup_{t\in[0,\tau]} \left| \frac{\hat{M}_{K_p}(t)}{\Phi_{K_{pp}}(t)} \right|$$

y su distribución estimada, la cual es el valor más grande observado en los residuales relativo a su varianza, denotado por la matriz $\Phi_K(t)$.
Los detalles sobre el desarrollo de los residuales de martingalas acumulativos, así como la estimación de su varianza, puede encontrarse explicado por *Martinusse & Scheike* (págs. 151, 228 y 260, 2006) para los modelos aditivos, multiplicativos y aditivo-multiplicativos respectivamente.

Capítulo 3

Ajuste de modelos de regresión

A continuación se presentan dos ejemplos de los modelos descritos en los capítulos anteriores. El primero de ellos se realiza con una base de datos que describe tiempos de infección de enfermedades venéreas. Con ella se ejemplifican los diferentes modelos explicados en el **capítulo 2** con el fin de mostrar los resultados que se pueden obtener al ajustar modelos con covariables con efecto constante en el tiempo β. Se realizan las pruebas de hipótesis para corroborar que efectivamente se tienen efectos constantes en el riesgo de infección. Los datos se encuentran en el paquete *KMsurv* de R nombrados como *std*, así como en la página web de los autores *Klein & Moeschberger*: *https://www.mcw.edu/Biostatistics.html*.

La segunda base de datos consiste en un grupo de pacientes con cirrosis biliar primaria a quienes se les estudia el riesgo de fallecimiento. Este análisis se lleva acabo para mostrar los alcances del análisis cuando se considera que las covariables no tienen un efecto constante en el riesgo. Esto debido a que en las pruebas de hipótesis realizadas para estudiar si el efecto de las variables era constante, produjeron evidencia para considerar que no. Esta base de datos se incluye en el paquete *survival* bajo el nombre de *pbc* y son presentados por *Fleming & Harrington* (1991, pág. 359). Sin más preámbulos, se procede a describir las bases de datos junto con el correspondiente análisis.

3.1. Estudio de enfermedades de transmisión sexual

Modelo de riesgos proporcionales

Con el fin de ejemplificar los modelos anteriores, se procede a hacer el análisis de los datos que fueron recogidos en un estudio de enfermedades venéreas realizado en varias poblaciones en EEUU donde es común la prevalencia de enfermedades de transmisión sexual. Por ello, se analizan datos recolectados en *Klein & Moeschberger* (2003, pág 13) correspondientes a 877 pacientes de sexo femenino las cuales habían presentado gonorrea o clamidia. Posteriormente

se midió el tiempo que pasó desde que se curaron de alguna de esas enfermedades hasta que se volvió a detectar el contagio de gonorrea o clamidia. Ambas enfermedades representan un riesgo latente en la población debido a dos factores principales:
Primero, son padecimientos asintomáticos, de tal modo que no se pueden percibir con facilidad los síntomas y a menudo se confunden con otros padecimientos de menor importancia. Y por otro lado, a pesar de que se tienen tratamientos con antibióticos para tratar y curar exitosamente ambas enfermedades, suceden muchos casos en los que se vuelve a presentar la infección y por lo tanto, se vuelve a poner en riesgo la salud de la población.
Dado que estas son enfermedades que son fácilmente prevenibles y tratables, es una incógnita el porqué persisten en las poblaciones, a tal grado de ser una epidemia en potencia. Los efectos de estas enfermedades pueden tener impactos más fuertes en las mujeres, por ello se recolectaron datos de mujeres únicamente con el fin de estudiar la prevalencia de las enfermedades en ellas.

El propósito de este estudio es identificar los factores de riesgo en los pacientes, los cuales provocan una reaparición de estas enfermedades. Y dado que se tiene la hipótesis de la posible existencia de un grupo poblacional en el que las enfermedades prevalezcan, se pretende también identificarlo para evitar que las enfermedades se sigan expandiendo.

Descripción de los datos

Las variables que se registraron en el estudio se describen en el cuadro **3.1** y se separan en 3 diferentes tipos: demográficas, conductuales y sintomáticas. Las variables demográficas incluyen la información referente al grupo poblacional en el que se incluye la paciente, así como antecedentes médicos y personales, las conductuales contienen información referente a los hábitos de la vida sexual de la paciente y las sintomáticas contienen información que describe síntomas físicos de la paciente, los cuales podrían asociarse a ambas enfermedades.
Vale la pena mencionar que no se tienen valores imputados.

Cuadro 3.1: Descripción de variables incluidas en la base de datos *std* del paquete *KMsurv* en R.

Nombre	Descripción	Codificación/Unidades
time	Es el tiempo hasta la reaparición de la infección o hasta la pérdida de seguimiento de la paciente.	Días
rinfct	Indica si se observó la reinfección. En otro caso se tiene censura por la derecha.	1: se presentó la infección de nuevo. 0: en otro caso.

Variables demográficas		
race	Color de piel de la paciente	1: para caucásica. 0: para negra.
age	Edad de la paciente	Años
yschool	Años de formación académica	Años
marital	Indica el estado civil del individuo	0: para casado. 1: para divorciado. 2: para soltero.
iinfct	Tipo de enfermedad adquirida	1: para gonorrea. 2: para clamidia. 3: para ambas.
Variables conductuales		
npartner	Número de parejas sexuales	Cantidad
os12m	Indica si se practicó sexo oral en los últimos 12 meses	1: se practicó sexo oral. 0: en otro caso.
os30d	Indica si se practicó sexo oral en los últimos 30 días	1: se practicó sexo oral. 0: en otro caso.
rs12m	Indica si se practicó sexo anal en los últimos 12 meses	1: se practicó sexo anal. 0: en otro caso.
rs30d	Indica si se practicó sexo anal en los últimos 30 días	1: se practicó sexo anal. 0: en otro caso.
condom	Indica la frecuencia con la que se usa condón	1: siempre se usa. 2: alguna veces se usa. 3: nunca se usa.
Variables Sintomáticas		
abdpain	Indica la presencia de dolor abdominal en el diagnóstico.	1: si se presenta dolor abdominal. 0: en otro caso.
dyscharge	Indica la señal de flujo vaginal.	1: si se presenta flujo vaginal. 0: en otro caso.
dysuria	Indica la presencia de disuria.	1: si se presenta disuria. 0: en otro caso.
itch	Indica la presencia de sarpullido.	1: si se presenta sarpullido. 0: en otro caso.
lesion	Indica si se tienen lesiones en el área.	1: si se presentó una lesión. 0: en otro caso.
rash	Indica la presencia de erupciones.	1: si se presentó una erupción. 0: en otro caso.
lymph	Indica señales de participación linfática.	1: si se presentaron señales. 0: en otro caso.
vagina	Indica si fue necesario analizar la vagina en la prueba médica.	1: si fue necesario. 0: en otro caso.
dchexam	Indica si hubo flujo vaginal durante la examinación.	1: si se presentó. 0: en otro caso.
abnode	Indica la presencia de algún nodo anormal.	1: si se presenta disuria. 0: en otro caso.

Análisis exploratorio de las variables.

Se muestran los datos correspondientes a la base *std*. Con el fin de ajustar los modelos, algunas variables se modificaron de la siguiente manera:

- *iinfct*: Para poder hacer un análisis más específico con relación a qué enfermedad se tenía previamente al entrar la paciente en el estudio, se va a dividir esta variable en las siguientes dos variables indicadoras:

  gono: Indicará 1 en el caso en que la paciente tenga gonorrea al inicio del estudio.
  clam: Indicará 1 en el caso en que la paciente tenga clamidia al inicio del estudio.

  En la **fig. 3.1** se observa el estimador de Kaplan-Meier estratificado por enfermedades. Se puede ver que el hecho de haber tenido el virus de la gonorrea previamente, aumenta considerablemente la probabilidad de verse una infección nuevamente.

- *marital*: En relación al estado civil de las pacientes, se considera que el hecho de que las pacientes no estén casadas, las hace igualmente propensas a una infección sin importar si son divorciadas o si son solteras. Esto se puede observar en el **cuadro 3.2**. Por ello se colapsan esas dos categorías y ahora sólo vale 1 si se está casado y 0 en otro caso.

- *age*: En la **fig. 3.2** se muestran curvas de Kaplan-Meier estratificadas para los cuartiles de la variable *age*. Lo que estas curvas muestran es que se tiene aproximadamente la misma tasa de supervivencia para personas que superan los 17 años, mientras que para aquellas pacientes que a lo más tienen 17 años, se tiene un incremento evidente en el riesgo de desarrollar reinfecciones. Por ello se decide colapsar esta variable a una variable indicadora:

  less17: Indica 1 si es menor de edad y 0 en otro caso

- *npartner*: Se tienen 70 casos que presentaron una reinfección a pesar de afirmar que no tenían parejas sexuales. Se ha demostrado que la infección se puede presentar fuera del coito, pero en muy pocos casos. Por lo tanto, se puede pensar en que al menos se tenía una pareja para estos casos realmente. Siendo así, se unen estos 70 casos a la categoría de los 607 de pacientes con una pareja sexual.

Cuadro 3.2: Casos de reinfección por cada grupo dependiendo el estado civil.

	rinfct		% de infección
marital	0	1	
0	19	9	47.36
1	44	16	36.36
2	467	322	68.85

Figura 3.1: Estimador de Kaplan-Meier estratificado por la variable *iinfc* de la base *std* para cada grupo de pacientes según las enfermedades previamente adquiridas (izquierda). Acercamiento del las mismas estimaciones para los primeros 250 días (derecha). En la parte inferior se muestra el número de pacientes en riesgo al tiempo indicado por cada estrato.

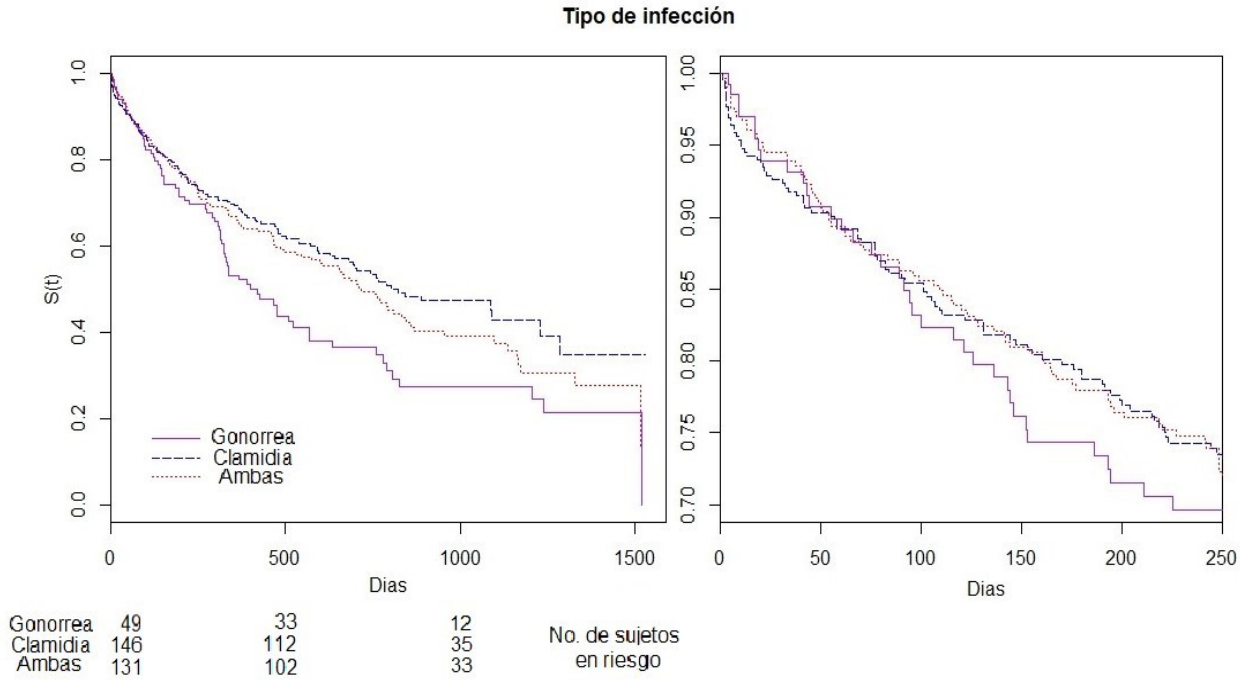
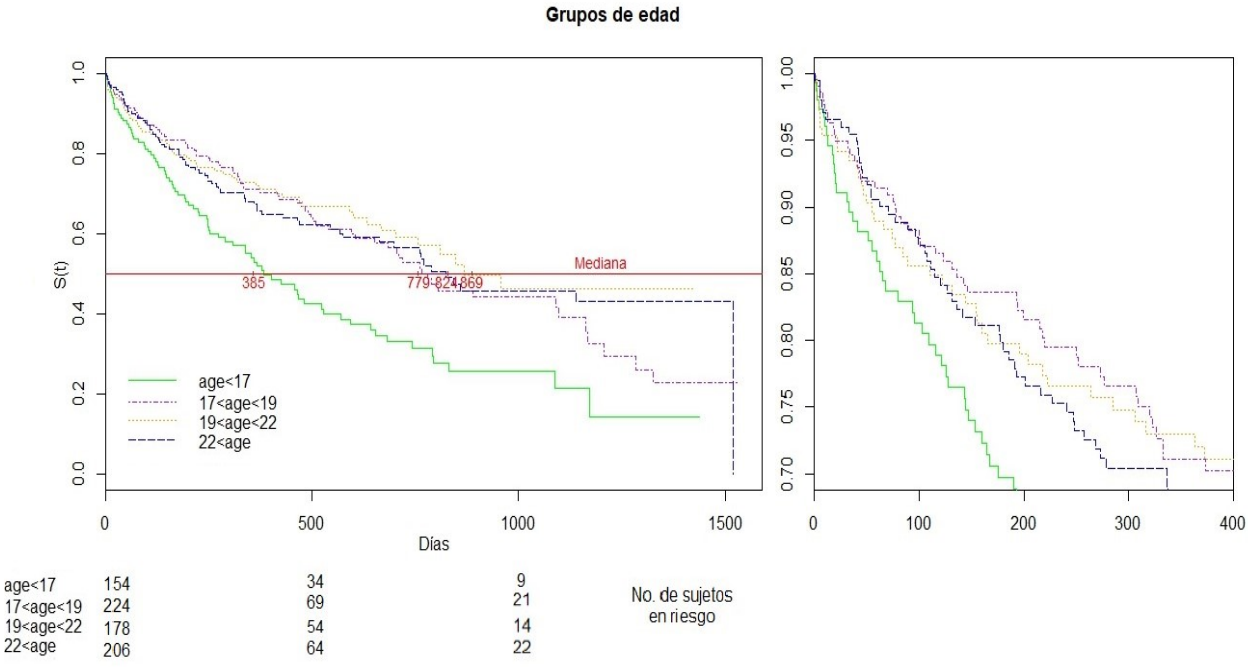
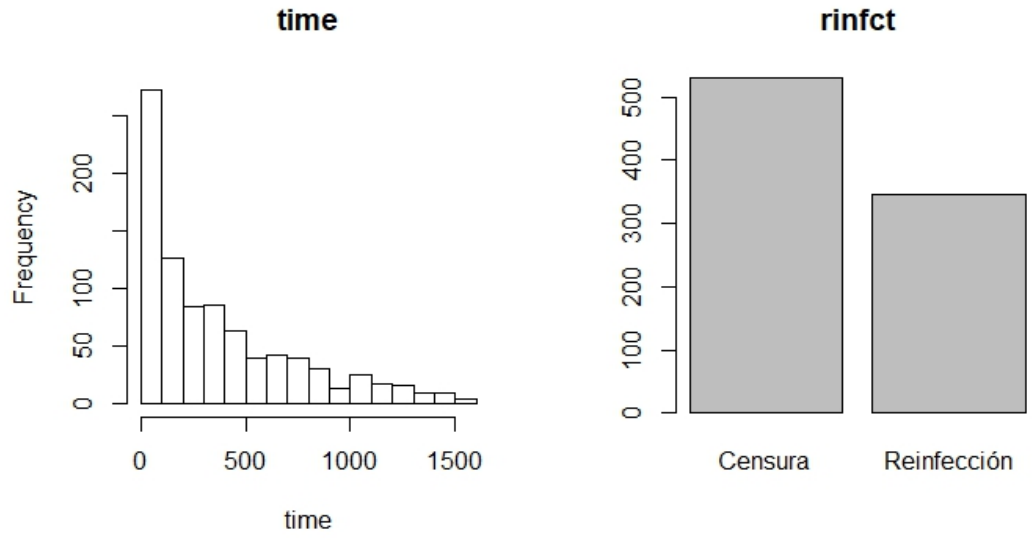


Figura 3.2: Estimador de Kaplan-Meier estratificado para cada grupo de edad definido según sus cuartiles. Incluye los valores de la mediana para cada grupo (izquierda). Acercamiento del las mismas estimaciones para los primeros 400 días (derecha). En la parte inferior se muestra el número de pacientes en riesgo al tiempo indicado por cada estrato.



Por otro lado, la variable con relación al tiempo de ocurrencia de la infección se distribuye de un modo inusual según la **fig. 3.3**, pues se observan muy pocos datos en los valores más altos. Las repercusiones de esta distribución de la variable *time* se observarían en las curvas realizadas para los coeficientes de regresión acumulados y pruebas de hipótesis. Más aún, puede ocasionar problemas de singularidad al hacer las estimaciones con las funciones del paquete *timereg*. La consola de R imprime el mensaje *"X'X not invertible at time t"*, con t los tiempos en los que se presenta singularidad. Esto se debe a que los saltos del proceso de conteo $N(t)$ serán menos frecuentes para valores grandes de t, lo cuál deriva en matrices $\boldsymbol{X}^T(t)\boldsymbol{X}(t)$ mal condicionadas (casi singulares) dentro de la estimaciones. Por ello se harán las estimaciones truncando las observaciones que superen los 900 días.

Figura 3.3: Distribución de la variable del tiempo de infección (izquierda) y número de caso de reinfecciones (derecha).



La distribución del resto de las variables se observan en las siguientes figuras. Se observa que casi toda la población se encuentra en un statu fuera de una relación (96.8 %).
Las variables de edad y años de formación académica (*age* y *yschool*) siguen una distribución unimodal, aunque *yschool* se comporta de un modo similar a la distribución normal. Estas variables tienen una moda de 26 años de edad y 12 años de formación académica respectivamente.

Respecto a la variable *npartner* del número de parejas sexuales, se puede ver en la **fig. 3.5** que se tienen algunos valores muy lejanos a los restantes que se acumulan en el valor de 1. Pueden representar outliers o puntos de influencia en los modelos que se ajustarán posteriormente.

Figura 3.4: Distribución de las variables de **a)** edad, **b)** raza, **c)** statu marital, **d)** años de formación académica y presencia de **e)** gonorrea y **f)** clamidia de la base *std*.

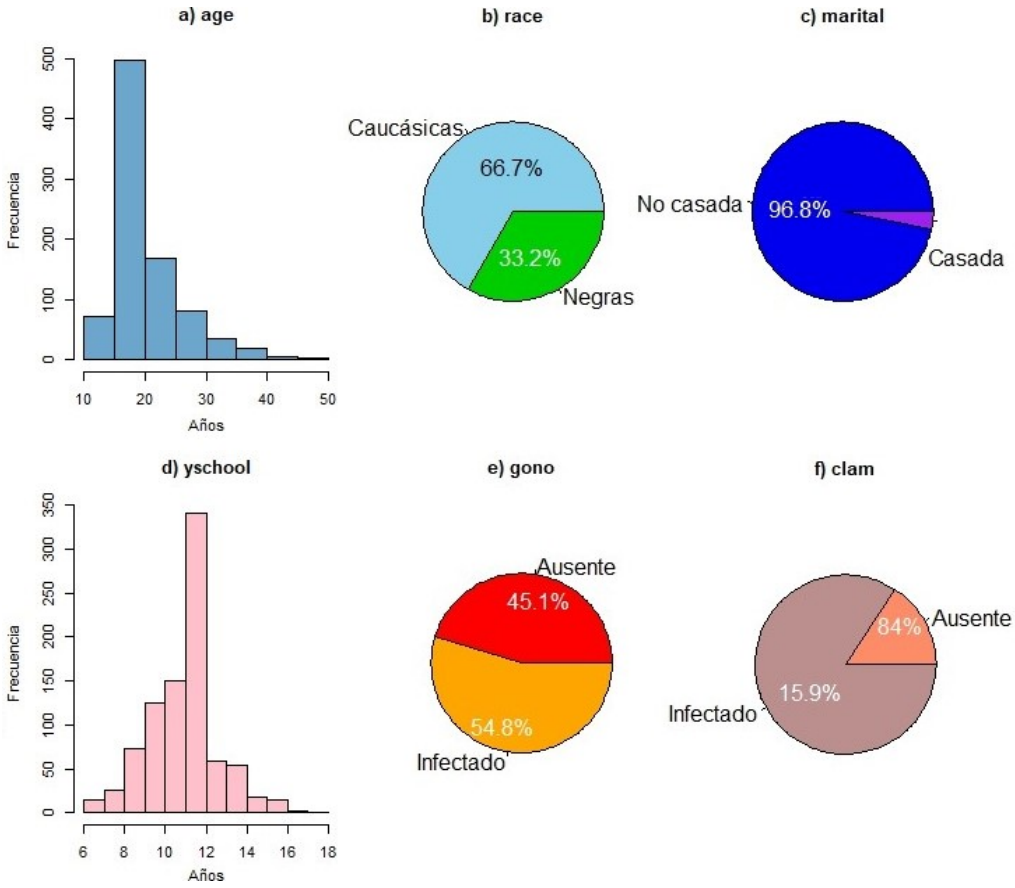
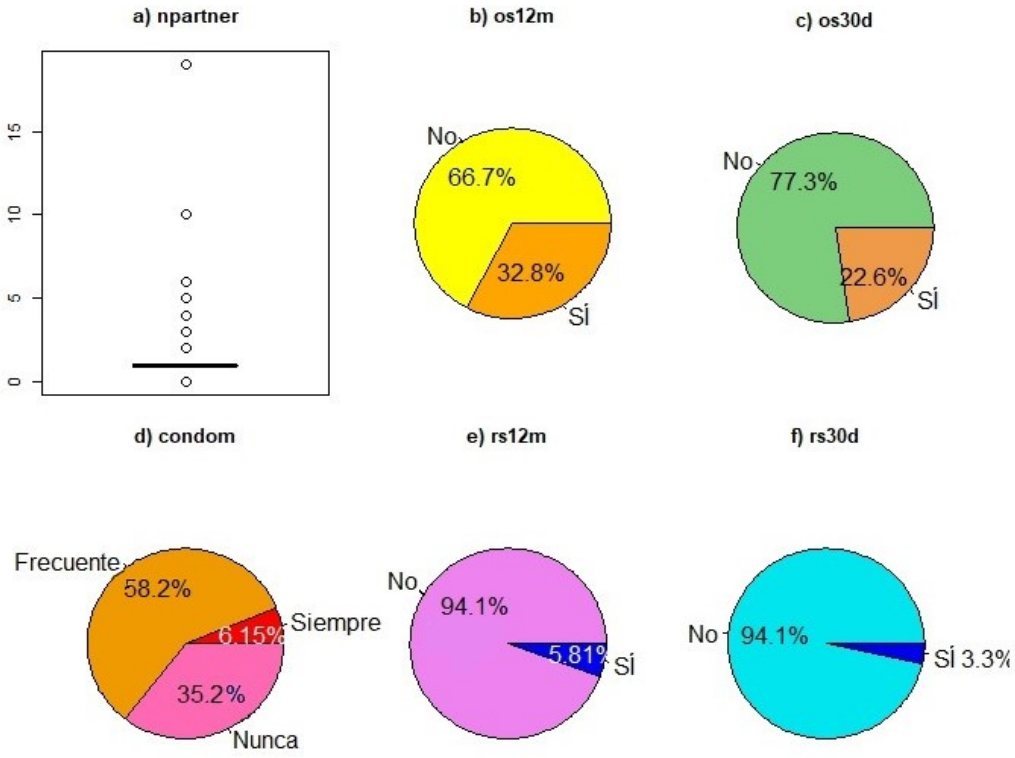
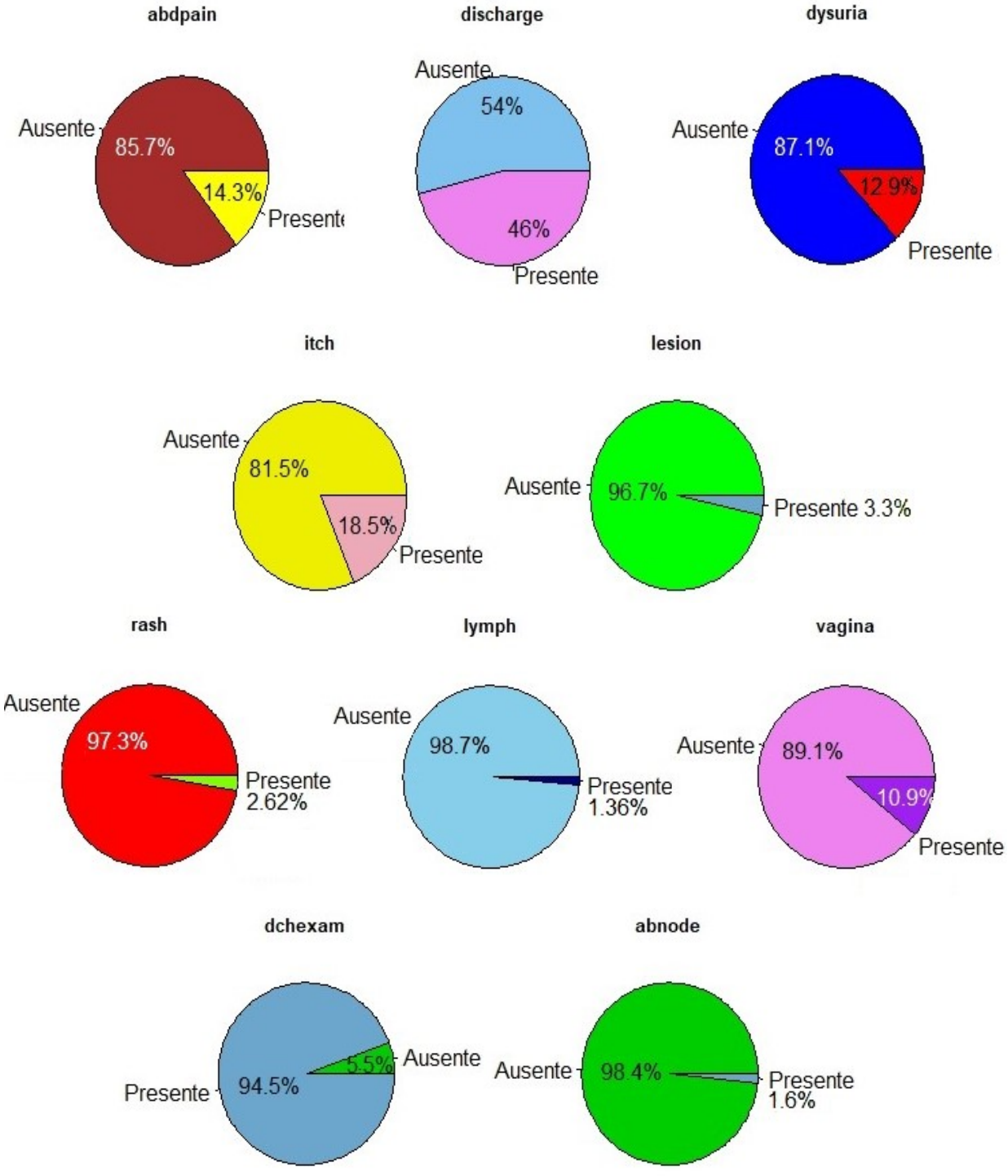


Figura 3.5: Distribución de las variables de **a)** número de parejas, tipo de relaciones (oral o rectal) en el último mes o año **b), c), e) y f)** y uso del condón **d)** de la base *std*.



Sobre las variables sintomáticas, sólo resaltan las distribuciones de las pacientes que habían presentado lesiones, sarpullido, nodos anormales y participación linfática al ser diagnosticadas , pues se registraron muy pocas pacientes que presentaron las características anteriores.

Figura 3.6: Distribución de las variables sintomáticas de la base *std*.



En el **cuadro. 3.3** se resumen la información de todas las variables incluidas en la base de datos expresadas en cuantiles y medias.

Cuadro 3.3: Resumen de los cuantiles y media de las 22 variables explicativas de la base *std* para las pacientes del estudio.

	Variables	*Min*	*1. C*	*Med*	*3. C*	*Max*	*Media*
Continuas	*age*	13	17	19	22	48	19
	yschool	6	10	12	12	18	11.43
	npartner	0	1	1	1	19	1.26
	Variables	*Valor 0*			*Valor 1*		
Categóricas	*race*	585			292		
	discharge	472			405		
	gono	396			481		
	chla	140			737		
	dysuria	763			114		
	marst	849			28		
	itch	714			163		
	os12m	589			288		
	lesion	848			29		
	os30d	678			199		
	rash	854			23		
	rs12m	826			51		
	lymph	865			12		
	rs30d	848			29		
	rash	781			96		
	abdpain	751			126		
	dchexam	48			829		
	abnod	863			14		
	condom	1: 64		2: 511		3: 312	

Selección de variables

Se decide elegir un conjunto de variables de considerable significado médico para investigar interacciones para más tarde ponderar el resultado de la función *stepAIC()*. Posteriormente, se determina investigar interacciones con respecto a las variables referentes a los años de formación académica *yschool* y la variable indicadora *less17* para las pacientes menores de edad. Esto se realiza dado que se quiere investigar si existe una interacción significativa entre los grupos de edad y los años de formación académica. De este modo se seleccionan todas las variables para el ajuste del modelo más grande. El ajuste del modelo de riesgos proporcionales inicial se muestra en el **cuadro 3.4 a)**. La función *stepAIC()* calcula el siguiente modelo reducido, cuyo resumen se muestra en el **cuadro 3.4 b)**.

$$\lambda(t) = \lambda_0(t) \cdot exp(\beta_1 \cdot chla + \beta_2 \cdot os12m + \beta_3 \cdot condom + \beta_4 \cdot npartner + \beta_5 \cdot less17 \cdot yschool) \tag{3.1.1}$$

Cuadro 3.4: Ajuste del modelo de riesgos proporcionales con las variables de la base *std* junto con su verosimilitud. Se muestran el modelo completo **(a)** y modelo simple seleccionado por algoritmo de la función *stepAIC()* del paquete *MASS* en el método *backwards*. **(b)**

a) *log verosim:*-2044.014	*Coef.* $\hat{\beta}_i$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
1. chla	-0.3181	-2.166	0.0303
2. os12m	-0.2659	-1.262	0.2070
3. condom	-0.2020	-2.078	0.0377
4. npartner	0.0943	1.826	0.0678
5. less17:yschool	0.0353	2.474	0.0134
6. marst	-0.3129	-0.868	0.3855
7. race	-0.1458	-1.042	0.2974
8. gono	0.0794	0.625	0.5318
9. os30d	-0.3403	-1.423	0.1546
10. rs12m	-0.0922	-0.208	0.8349
11. rs30d	0.0040	0.007	0.9942
12. abdpain	0.2190	1.434	0.1515
13. discharge	0.1609	1.429	0.1529
14. dysuria	0.1589	1.025	0.3054
15. itch	-0.1380	-0.899	0.3686
16. lesion	-0.0118	-0.036	0.9712
17. rash	0.0821	0.209	0.8347
18. lymph	-0.0756	-0.139	0.8893
19. dchexam	-0.4469	-1.951	0.0511
20. abnode	0.1096	0.256	0.7978
b) *log verosim:*-2050.333	*Coef.* $\hat{\beta}_i$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
1. chla	-0.3736	-2.823	0.0047
2. os12m	-0.5643	-4.223	$< 10^{-4}$
3. condom	-0.2046	-2.137	0.0325
4. npartner	0.1058	2.132	0.0329
5. less17:yschool	0.0355	2.525	0.0115

Para observar si realmente el modelo completo con todas las variables realiza un mejor ajuste que el modelo simple **3.1.1**, se contrasta la siguiente hipótesis

$$H_0 : \beta_j = 0 \; \forall j = 6, ..., 20 \;\text{ vs. }\; H_1 : \beta_j \neq 0 \text{ para alguna } j = 6, ..., 20$$

mediante el cociente de verosimilitud. Si el ajuste resulta no ser significativamente distinto al modelo simple, se optará por elegir el modelo con menos variables. Es valor del cociente es:

$$-2(\log(\mathscr{L}(reducido) - \log(\mathscr{L}(completo)) = -2(-2050.333 + 2044.014) = 12.63$$

el cual sigue una distribución asintótica $\chi^2_{(15)}$. Dado que el cuantil al 95 % de la distribución es 21.02, se tiene que el modelo completo no se ajusta significativamente mejor a los datos que el modelo más simple. Por lo tanto, se prestará atención al modelo en el **cuadro 3.4 b)**.
Como comentarios acerca de este modelo, hay que señalar que no se tiene evidencia estadística de un impacto en el diagnóstico por parte del factor de sexo rectal ni del hecho de haber

desarrollado gonorrea previamente, pues los coeficientes asociados a estas variables no son significativos. También hay que destacar el impacto que tiene la clamidia en el diagnóstico, pues efectivamente muestra una disminución en el riesgo de presentar una reinfección, siendo su coeficiente negativo. Respecto a la interacción de los años de estudio y el ser menor de edad, se tiene que aquellas pacientes menores de 17 años tienen mayor riesgo de adquirir nuevamente la infección durante su vida académica, pues el coeficinte asociado es positivo. Y el efecto de haber tenido sexo oral en un largo plazo descrito por *os12m* también parece disminuir el riesgo.

Para corroborar la hipótesis de correlación entre variables posteriormente, en el **cuadro 3.5** se muestra la matriz de correlación entre las variables seleccionadas en el modelo **3.1.1**. Hay que prestar especial atención sobre la variable *npartner*, pues a continuación se presenta un análisis para considerar alguna transformación para esta variable en el modelo.

Cuadro 3.5: Matriz de correlaciones para las variables del modelo **3.1.1**

	chla	*os12m*	*npartner*	*condom*	*less17:yschool*
chla	1.00				
os12m	0.05	1.00			
npartner	-0.03	0.17	1.00		
condom	-0.03	0.04	-0.12	1.00	
less17:yschool	-0.03	-0.17	-0.04	-0.09	1.00

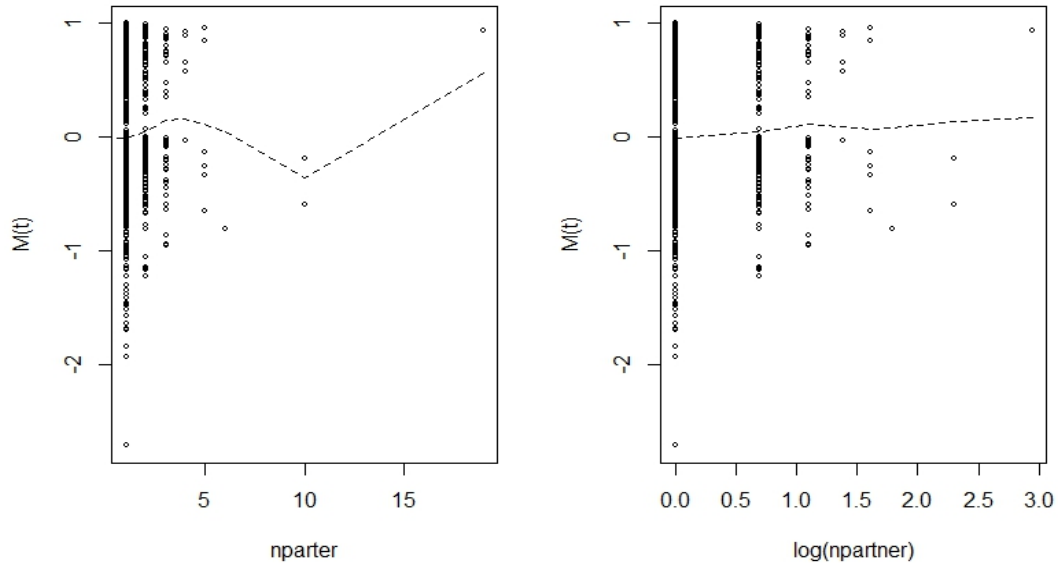
Evaluación de transformaciones

Con el fin de evaluar el uso de posibles transformaciones, se procede a analizar los residuales de las variables continuas del modelo. Como se muestra en la **sección 2.5**, para analizar la aplicación de una posible transformación, se puede estimar la función adecuada en la gráfica de los residuales de martingala del modelo (en el que no se incluye la variable en cuestión) vs. los valores de la variable en cuestión.

En la **fig. 3.7** se muestran los residuales de martingala del modelo con las variables de *chla*, *os12m*, *condom* y la interacción *less17:yschool* graficadas vs. *nparter* (izquierda) y *log(npartner)* (derecha). En la gráfica del lado izquierdo se muestra que sin transformar la variable *nparter*, el comportamiento de los residuales se concentra en los primeros valores. Lo cuál hace más notoria la influencia de los datos a la mitad del rango de *nparter*. Por otro lado, se decide hacer lo mismo para otra gráfica con los residuales del mismo modelo, pero con la transformación *log(npartner)*.

Se muestra una línea de ajuste más horizontal. Indicando que esta transformación posee potencial para predecir mejor el riesgo. Para corroborar esto, se puede observar en el **cuadro 3.5** que la variable *npartner* no está fuertemente correlacionada con el resto de las variables en el modelo. Para el resto de las variables no se tiene evidencia para la inclusión de otra posible transformación.

Figura 3.7: Gráficas de los residuales de martingala del modelo con *clam*, *os12m*, *condom* y *less17:yschool* vs. *nparter* (izquierda) y vs. *log(npartner)* de la base *std*.



Por lo anterior se ajusta el siguiente modelo:

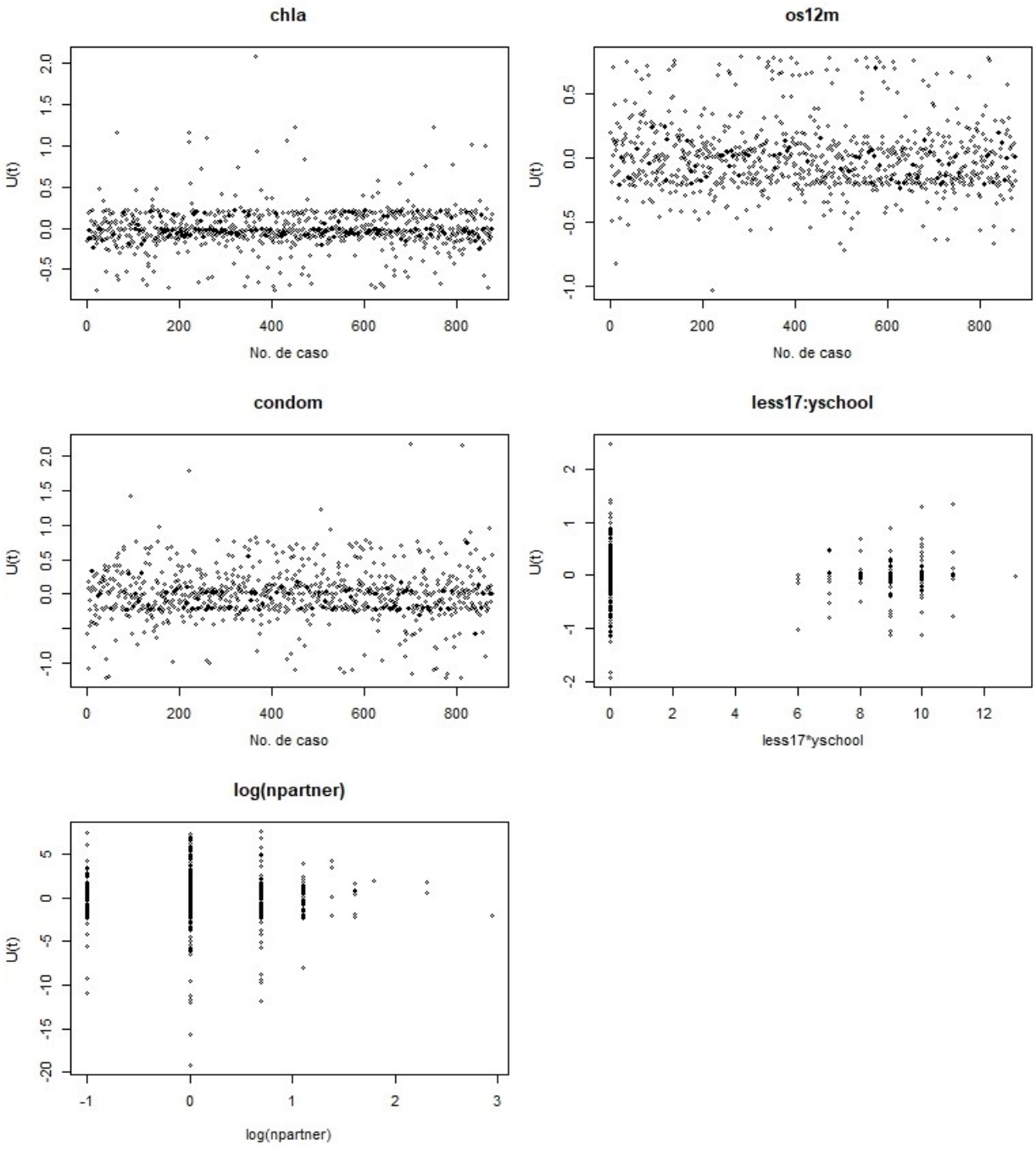
$$\begin{aligned}\lambda(t) = \lambda_0(t) \cdot exp(\beta_1 \cdot chla + \beta_2 \cdot os12m + \beta_3 \cdot condom + \\ + \beta_4 \cdot \log(npartner) + \beta_5 \cdot less17 \cdot yschool)\end{aligned} \tag{3.1.2}$$

Búsqueda de puntos de influencia

Como parte del análisis de residuales, se plantea la búsqueda de puntos de influencia, outliers o errores de medición con residuales de score. Como se sabe de la **sección 2.5**, los residuales de score, nos pueden ayudar a identificar los puntos que ocasionan una desviación importante del ajuste con respecto a los datos. En la **fig. 3.8** se pueden observar gráficas de los residuales de score de cada variable del modelo **3.1.1** vs. cada uno de los 877 casos. Para las variable continuas *less17:yschool* y *log(nparter)*, es adecuado graficar sus correspondientes residuales vs. sus respectivos valores.

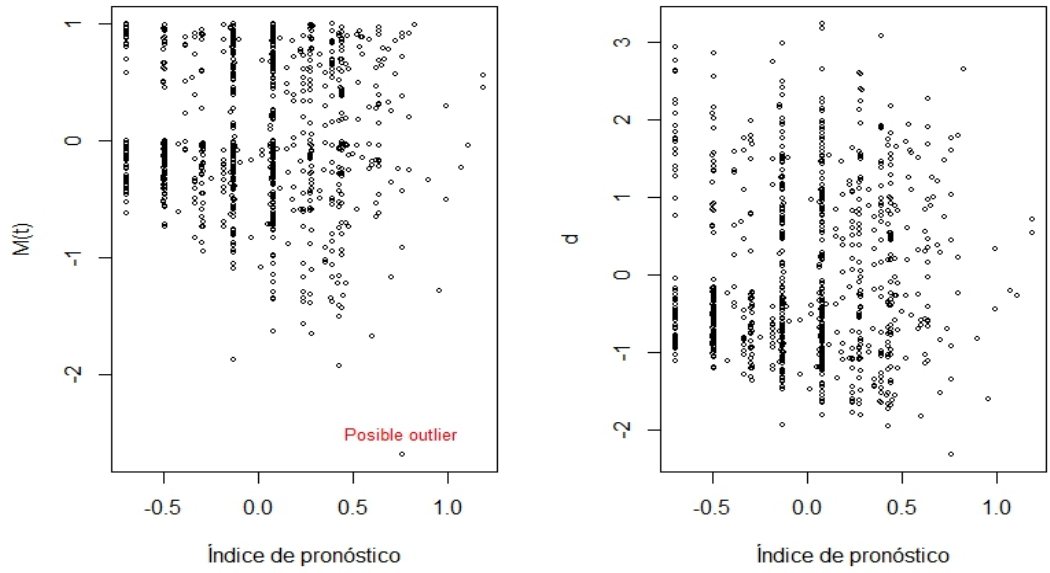
Se puede ver que hay algunos puntos de influencia en las gráficas correspondientes a las variables *condom* y *chla*, pero casos evidentemente apartados del resto (outliers) no son fácil de identificar en estas gráficas. Incluso para los valores cuyos residuos se apartan claramente del resto, se observa que no son fáciles de considerar como casos extraños u outliers, dada que la información de las variables no permite dar un diagnóstico que respalden esas afirmaciones. Aunque en general, salvo algunos puntos, nuestro modelo se comporta de un modo adecuado con respecto a los valores observados. Sin embargo, también es importante considerar los residuos de devianza para ver si estos mismos outliers se presentan en sus gráficas.

Figura 3.8: Residuales de score del modelo **3.1.4** para cada variable vs. número de caso (para *clam*, *os12m* y *condom*) y vs. valor correspondiente (*less17:yschool* y *log(nparter)*).



Como se sabe, los residuales de martingala también nos pueden aportar información acerca del modelo. Sin embargo, dado que no son simétricos y su rango de valores es $[0, \infty)$ es difícil detectarlos. Por ello, en la **sección 2.5** se describe la transformación de los residuales de martingalas dada por la definición de los residuales de devianza. En la **fig. 3.9** se muestran los residuales del modelo graficados vs. el índice de pronóstico, como otro apoyo gráfico para identificar outliers o valores atípicos.

Figura 3.9: Residuales de martingala (izquierda) y Residuales de devianza (derecha) del modelo **3.1.4** vs. Índice de pronóstico.

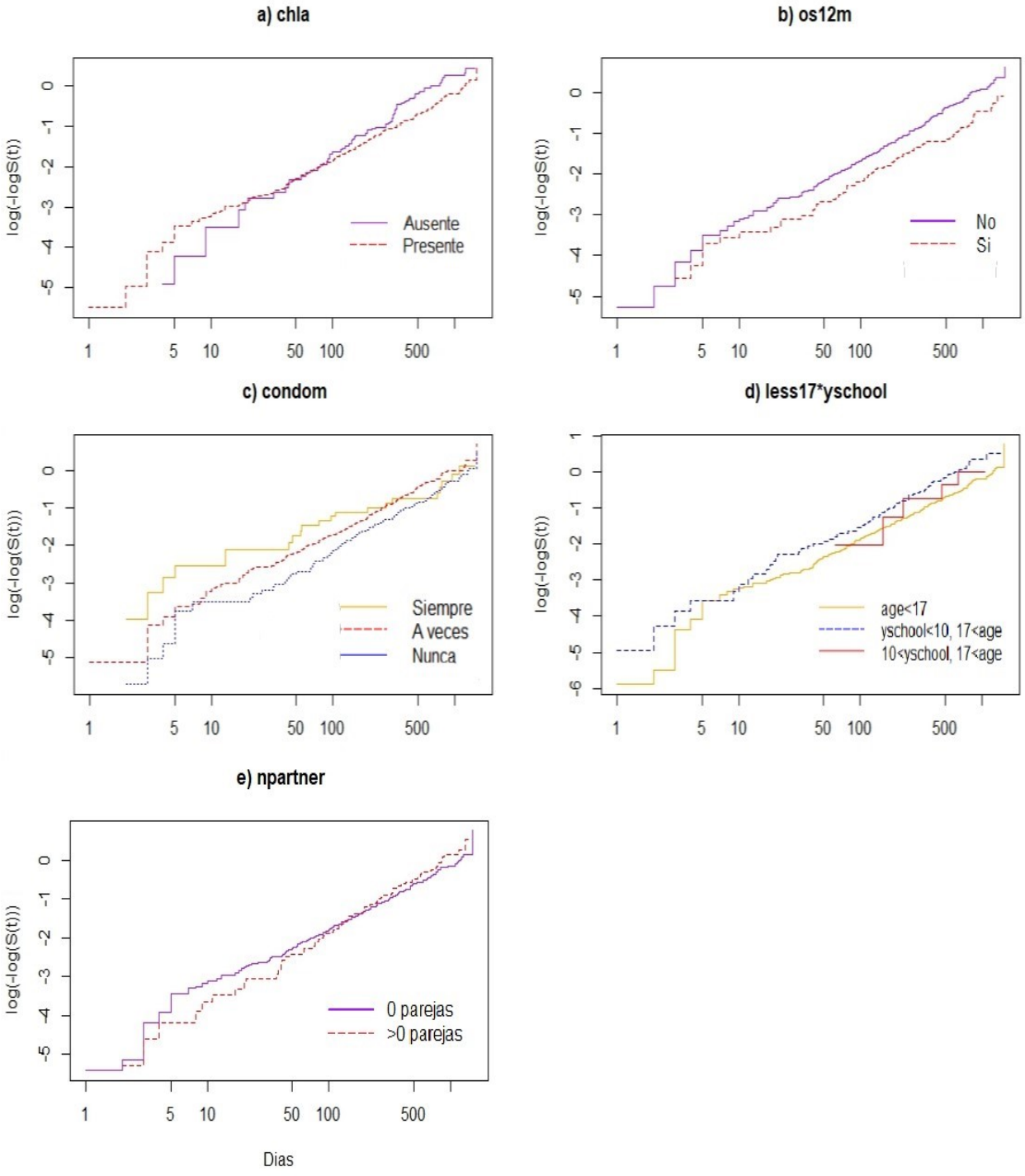


El único valor que podría identificarse como un outlier en la gráfica de residuales de martingala (izquierda), se observa más próximo a la tendencia del resto de los datos en los residuales de devianza (derecha). Por ello, podemos decir que si bien, algunos datos son puntos de influencia, no son outliers. Aunque no se descartan errores de medición.

Validación del supuesto de proporcionalidad

Finalmente, para considerar nuestro modelo como adecuado. Hace falta observar si efectivamente las variables tienen un efecto proporcional dentro de nuestro modelo. Para ello, se procede a elaborar estratificaciones en cuartiles para las variables continuas, con el fin de ver si las curvas estratificadas $log(-log(\hat{S}(t))$ son paralelas. En la **fig. 3.10** se muestra el resumen de los resultados.

Figura 3.10: Curvas estratificadas de $log(-log(\hat{S}(t))$ para cada variable del modelo **3.1.4**. Cada estimación se realiza con el estimador de Kaplan-Meier



Como se puede ver, hay algunas variables que podrían no cumplir con el supuesto de proporcionalidad. En la gráfica de la **fig. 3.10** se puede observar cierta evidencia para rechazar la hipótesis de proporcionalidad para las variables *chla* y *condom*. Sin embargo, las pruebas de proporcionalidad ejecutadas por la función *cox.zph()* muestran que son proporcionales a un nivel de 95 % de confianza. A continuación, en el **cuadro 3.6** se muestra la prueba para el resto de las covariables.

Cuadro 3.6: Prueba de proporcionalidad del modelo **3.1.2** para cada covariable y para el modelo en general. La columna *chisq* representa el valor de la estadística de la expresión **2.2.8** descrita en la **sección 2.2**, que se distribuye asintoticamente como $\chi^2_{(5)}$.

	T	*P-Valor*
chla	-0.03	0.63
os12m	0.01	0.92
condom	0.06	0.26
log(*partner*)	0.03	0.56
less17:yschool	0.50	0.48
GLOBAL		0.83

Lo que se puede determinar, es que bajo un nivel de confianza del 95 % las variables son proporcionales al riesgo de reinfección. Sin embargo, es importante considerar ajustar otros modelos en busca de un mejor ajuste.

Ajuste del modelo de riesgos proporcionales no paramétrico

El modelo anterior tiene la cualidad de describir de modo simple el efecto de las covariables sobre el diagnóstico de los pacientes. Sin embargo, en ocasiones es bastante evidente que algunas variables de nuestra base de datos *std* no siempre tienen el mismo efecto sobre el riesgo al que se someten los pacientes. Pues para distintos tiempos este efecto puede ser distinto. Por ello, el análisis anterior ha sugerido identificar cuáles son aquellas variables, si es que hay, que no siempre tienen el mismo efecto en el riesgo de reinfección en las pacientes. Con este motivo, se procede a ajustar un modelo de riesgos proporcionales no paramétrico.

Ajuste del modelo

Se ajusta el mismo modelo de la expresión **3.1.2** con la función *timecox()* de la paquetería *timereg* en R. Además de la hipótesis de significancia de los coeficientes:

$$H_0 : \beta_j(t) = 0 \text{ vs. } H_1 : \beta_j(t) = \beta_j$$

se desea entonces, probar la hipótesis de que las variables tienen un efecto constante a lo largo del tiempo, i.e. se desea investigar para cada una de las p variables, sus correspondientes coeficientes de regresión acumulados. En otras palabras, se contrasta la siguiente hipótesis para una constante γ:

$$H_0 : B_j(t) = \gamma t \text{ vs. } H_1 : B_j(t) \neq \gamma t$$

Para eso, se tienen las pruebas formales de Kolmogorov-Smirnov y el *Score process* descrito en el **capítulo 2**. La prueba de Kolmogorov-Smirnov se realiza automáticamente con la función de ajuste *timecox()* de la paquetería *timereg*, cuyos resultados son los siguientes:

Cuadro 3.7: Prueba de hipótesis de significancia de las variables y prueba de Kolmogorov-Smirnov para contrastar la hipótesis de efecto invariante del modelo **3.1.2**.

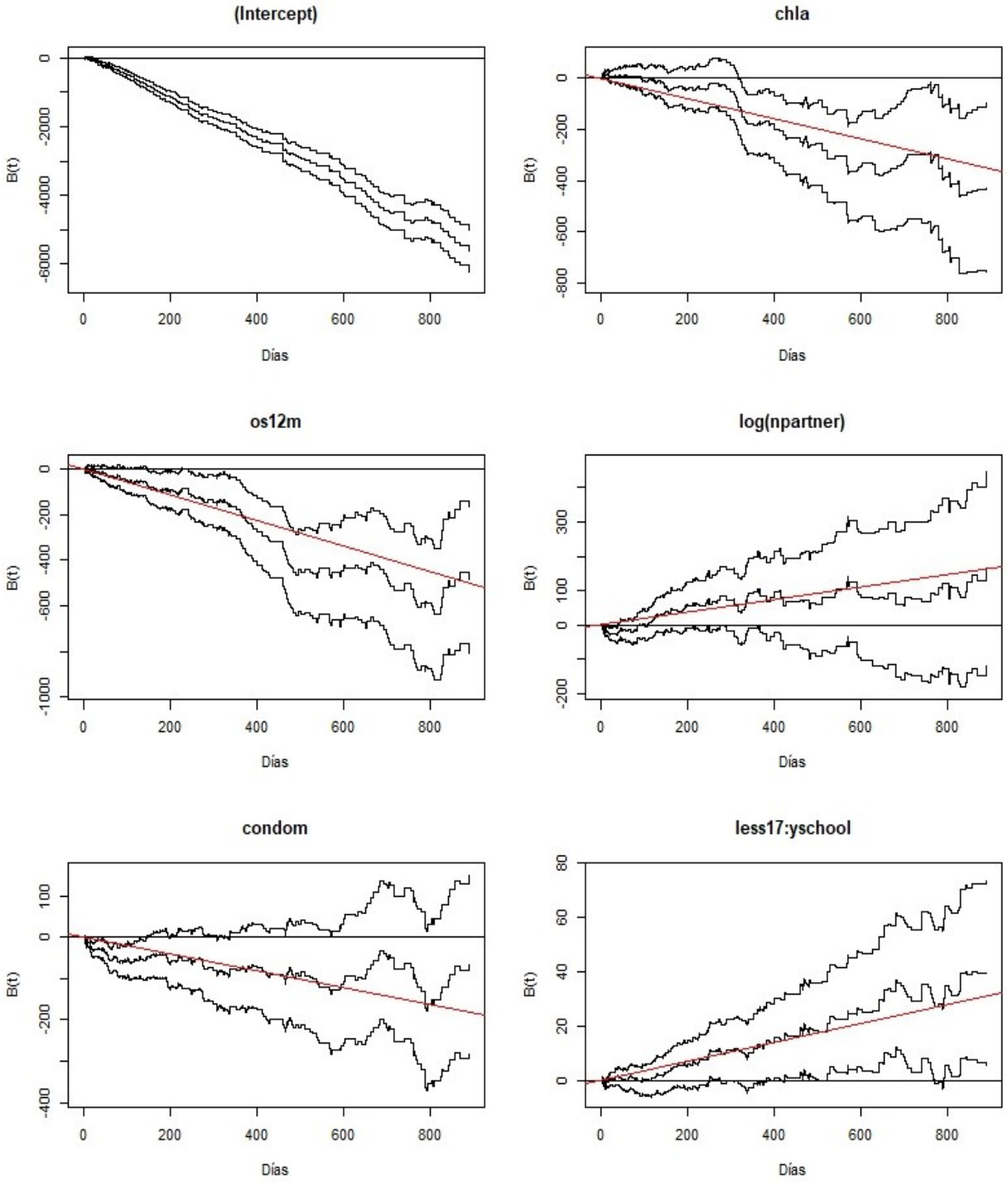
	$sup\|B(t)/\Phi(t)\|$	*P-Valor*	*Kolmogorov*	*P-valor*
intercepto	19	3×10^{-4}	338.0	0.253
chla	4.10	0.001	110	0.726
os12m	5.76	10^{-5}	110	0.726
log(*npartner*)	2.94	0.066	93	0.844
condom	2.99	0.061	118	0.228
less17:yschool	2.87	0.077	8.28	0.923

Se ajusta un modelo no paramétrico, dado por la siguiente expresión:

$$\begin{aligned}\lambda(t) = \lambda_0(t) \cdot exp(\beta_1(t) \cdot chla + \beta_2(t) \cdot os12m + \beta_3(t) \cdot \log(npartner) + \\ + \beta_4(t) \cdot condom + \beta_5(t) \cdot less17 \cdot yschool)\end{aligned} \tag{3.1.3}$$

En el **cuadro 3.7** se puede observar acorde con los p-valores, que las variables son significativas (bajo un nivel de significancia de 10 %). Adicionalmente, se muestra la prueba de Kolmogorov-Smirnov, la cual corrobora que no se puede rechazar la hipótesis de efecto constante.

Figura 3.11: Estimaciones de los coeficientes acumulados $\hat{B}(t)$ para el intercepto y todas las variables del modelo **3.1.3**. Las pendientes de las líneas rojas son los coeficientes estimados para los modelos paramétrico del **cuadro 3.8**.



En la **fig. 3.11** se observa una gráfica del coeficiente de regresión acumulado $B(t)$ vs. t de todas las variables del modelo en el **cuadro 3.7**, junto con bandas de confianza. Lo que se observa son coeficientes que indican que el efecto de las variables es constante a lo largo del tiempo. Esto porque las curvas se asemejan a pendientes cuyas derivadas son constantes.
Las únicas variables cuyas curvas parecen tener una pendiente no constante son *os12m* y *condom*, con algunos intervalos de tiempo donde se presentan variaciones importantes. Por otro lado, la significancia de las variables también se puede corroborar en las gráficas de coeficientes acumulados, pues si las bandas de confianza contienen la función $B(t) = 0$, se tiene cierta evidencia de que la variable podría no ser significativa.

Cuadro 3.8: Coeficientes estimados del modelo proporcional **3.1.1** paramétrico y su significancia.

	coef. $\hat{\beta}_i$	$\hat{\beta}_i/\sigma_i$	*P-Valor*
chla	-0.391	-2.90	0.004
os12m	-0.570	-4.14	8×10^{-4}
log(*npartner*)	0.282	1.82	0.069
condom	-0.196	-1.94	0.052
less17:yschool	0.034	2.31	0.021

Una vez que se ha determinado que significativamente, las variables tienen un efecto constante en el riesgo, se puede ajustar un modelo de riesgos proporcionales paramétrico. Sin embargo, los procesos de conteo ofrecen pruebas de hipótesis diferentes para dar más evidencia acerca de la exactitud del modelo y así corroborar más certeramente las hipótesis. Ajustando el modelo de riesgos proporcionales paramétrico con la función *timecox()* del paquete *timereg* se obtienen las estimaciones de los coeficientes del modelo paramétrico **3.1.2**, los cuales se asemejan mucho a los obtenidos en el **cuadro 3.4 b)** del modelo de riesgos proporcionales paramétrico.

Prueba de Proporcionalidad *(Score process)*

Mediante distribuciones aproximadas, se pueden simular curvas dada la hipótesis de proporcionalidad en el modelo y comparar estas curvas con la originada por el modelo con los datos especificados de la base *std*. La función *plot.timecox()* del paquete *timereg* muestra los resultados de las simulaciones del proceso bajo la hipótesis de proporcionalidad [1] y la curva de la función score evaluada en cada tiempo t del intervalo $[0, 900]$ de observación.

De las gráficas de la **fig. 3.12** se puede concluir que dado que sólo la función score de la variable *chla* se aparta levemente de las simulaciones, se tiene evidencia para considerar que esta variable no es proporcional. Más aún, el resumen del ajuste de la función *timecox()* muestra una prueba de proporcionalidad basada en el $sup|U_j(\hat{\beta}, t)|$ de la función score para la j-ésima variable, cuya distribución es aproximada. De tal modo que el **cuadro 3.9** contiene los resultados de la prueba

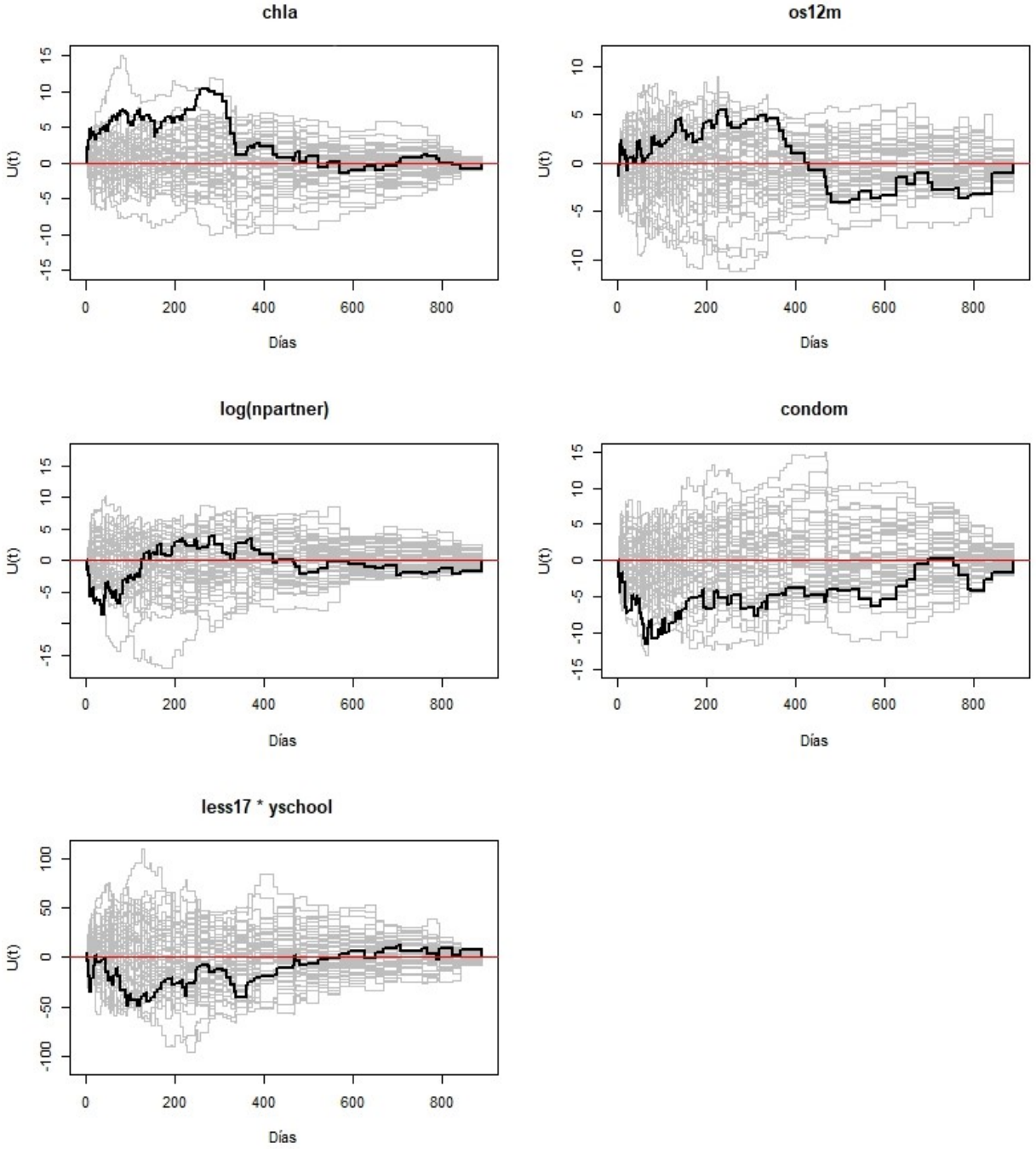
[1] Por defecto muestra 50 simulaciones.

de hipótesis siguiente sobre la función score:

$$H_0 : U_j(\hat{\beta}, t) = 0 \text{ vs. } H_1 : U_j(\hat{\beta}, t) \neq 0$$

Al respecto, se puede decir que las variables componen un modelo de riesgos proporcionales, aunque debido a la ligera desviación de la variable **chla**, quizás sea más apropiado incluirla en el análisis de un modelo de forma distinta.

Figura 3.12: Función score y 50 simulaciones para cada variable para contrastar las discrepancias de los datos con el modelo proporcional **3.1.3**.



Prueba de bondad de ajuste

Para evaluar el ajuste del modelo, se pueden analizar las curvas en los residuales de martingala acumulativos. En la **fig. 3.13** se muestran dichos residuales, descritos en la **sección 2.5**, para la variable *chla*, que aunque no cumple fuertemente con el supuesto de proporcionalidad, parece poder predecir adecuadamente el modelo dado que el proceso parece oscilar no lejos del 0.

Cuadro 3.9: Prueba de proporcionalidad para el modelo **3.1.3** dada por el proceso de scores.

	$sup\|U(t)\|$	*P-Valor*
chla	10.40	0.022
os12m	5.53	0.550
log(*npartner*)	6.69	0.286
condom	11.50	0.122
less17:yschool	49.50	0.598

Por otro lado, en la **fig. 3.14** y **fig. 3.15** se muestran los residuales estratificados para 0 y 1 parejas, y para 2 o más parejas (esto debido a su distribución concentrada alrededor del valor 1) correspondientes a la variable *nparents* con y sin la transformación de *log()* respectivamente. Dado que no existe gran diferencia entre ambos ajustes, se conserva la transformación *log()* en el modelo **3.1.2**.

Figura 3.13: Residuales de martingala acumulativos para *chla* del modelo **3.1.3**.

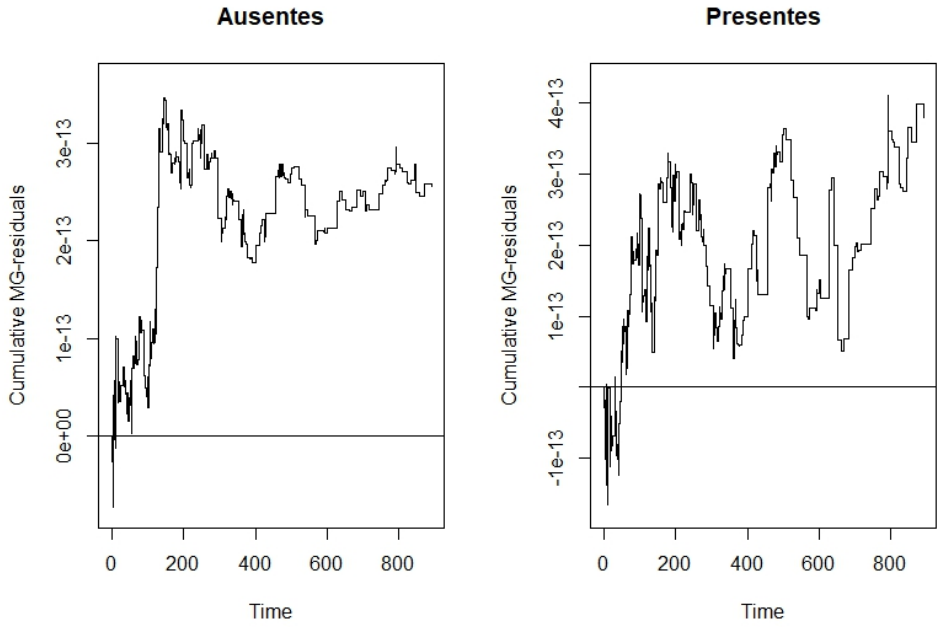


Figura 3.14: Residuales de martingala acumulativos de *log(npartner)* estratificados para 0 y 1 parejas (escasas) y para 2 o más parejas (varias) del modelo **3.1.3**.

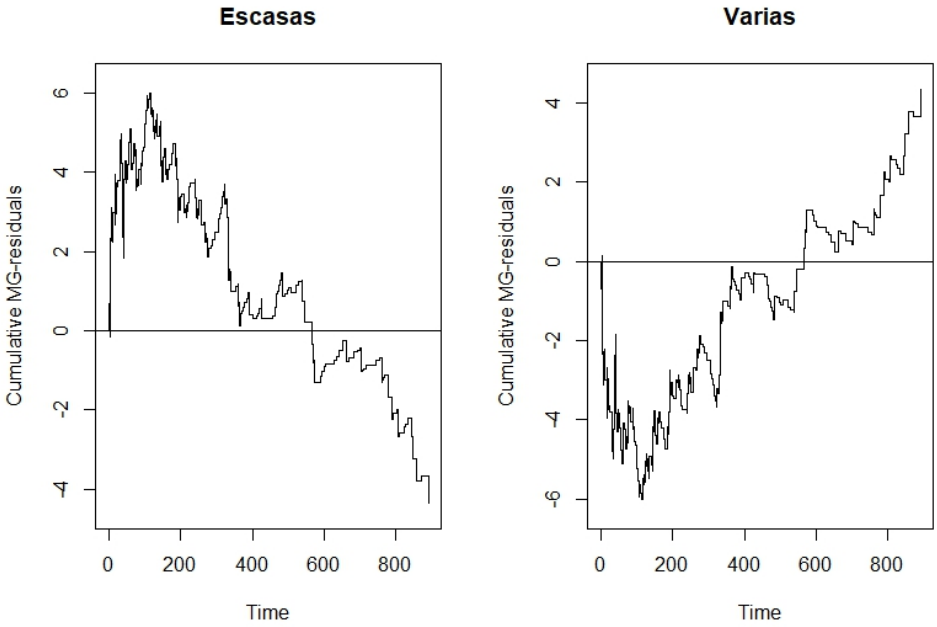
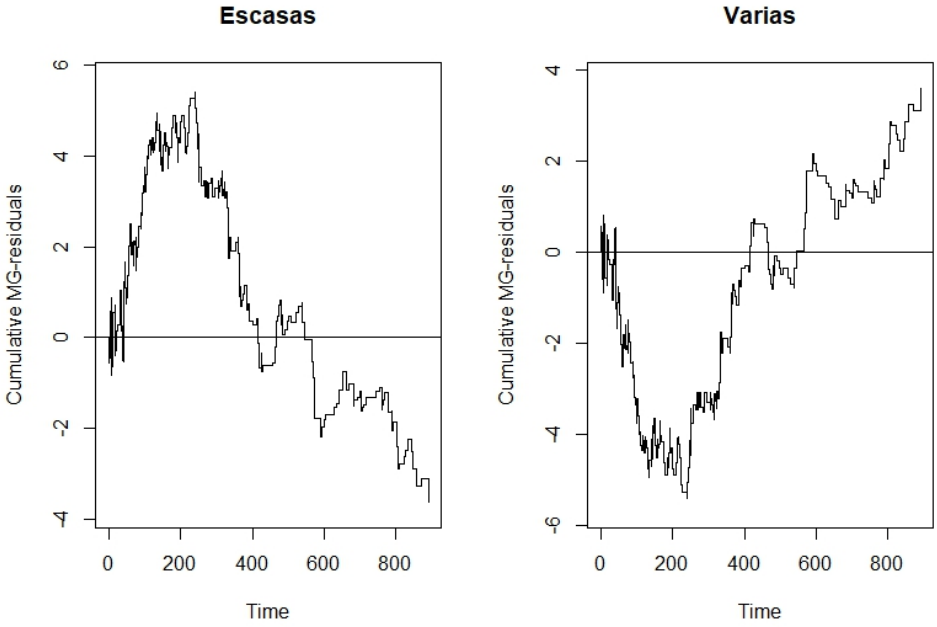


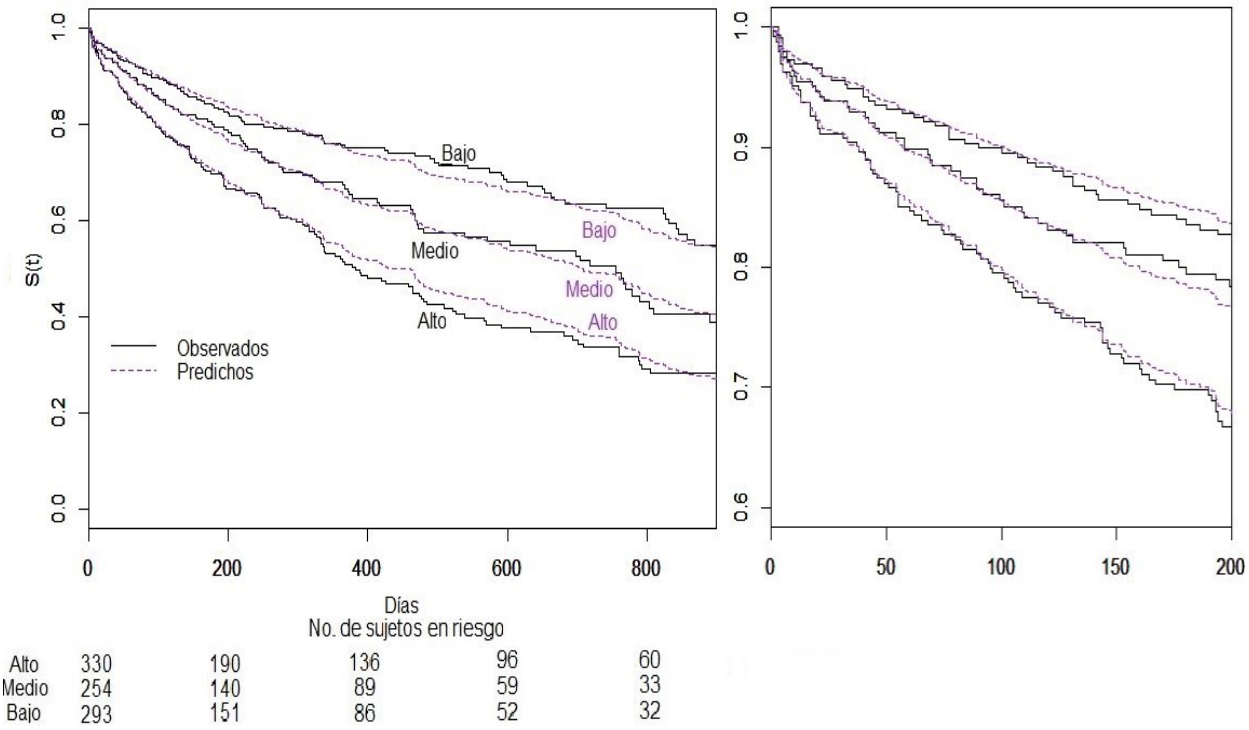
Figura 3.15: Residuales de martingala acumulativos de *npartner* estratificados para 0 y 1 parejas (escasas) y para 2 o más parejas (varias) del modelo **3.1.3**.



Grupos de riesgo

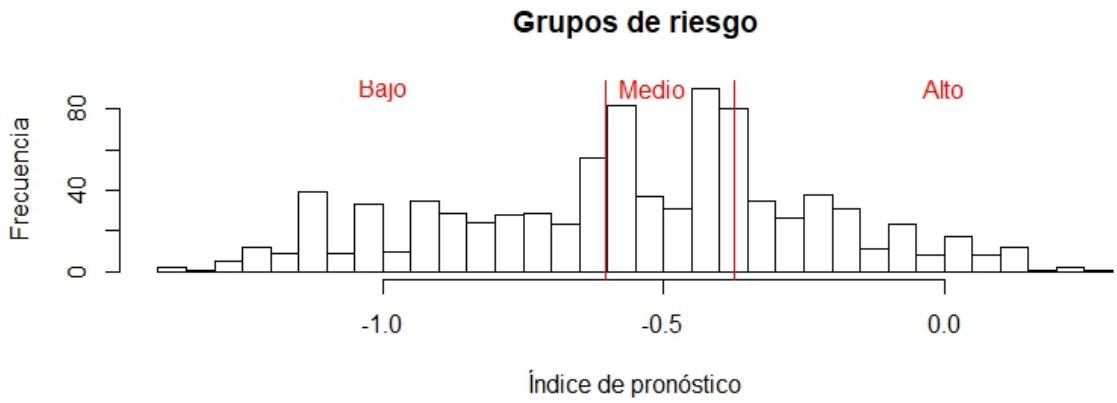
Dado que los coeficientes obtenidos en ambos ajustes son aproximadamente los mismos (según los cuadros **3.4 b)** y **3.8**), se opta por realizar el ajuste con el modelo de riesgos proporcionales obtenido con la paquetería *survival*. La clasificación de los grupos de riesgo se construyen a partir de los terciles del índice de pronóstico elaborado con los coeficientes estimados del modelo paramétrico **3.1.2**. Se realiza la estimación de Kaplan-Meier estratificada por los grupos elaborados y se compara, representando las observaciones, con la función de supervivencia estimada por el modelo evaluada en los valores medio de cada variable restringida a su grupo de riesgo correspondiente, lo cuál representa la predicción a comparar. Las curvas obtenidas se muestran en la **fig. 3.16**

Figura 3.16: Curvas con los grupos de riesgo (izquierda). Las observaciones están dadas por las curvas estratificadas de Kaplan-Meier y las predicciones están dadas por la media de cada variable restringida a cada grupo de riesgo, determinado por los terciles el índice de pronóstico del modelo **3.1.2**. Un aumento de la misma gráfica en los primeros valores se muestra a la derecha.



La predicción es bastante favorable con respecto a las observaciones y se tiene una buena clasificación para predicciones. La prueba log rank arroja un *p-valor* de 2.99×10^{-9}. Lo cual quiere decir que los grupos de riesgo definidos tienen funciones de supervivencia bien diferidas una de las otras.
La distribución del índice de pronóstico, junto con la separación de los valores correspondientes a cada grupo de riesgo, se muestra en la **fig. 3.17**.

Figura 3.17: Distribución del índice de pronóstico y grupos de riesgo.



El propósito en la siguientes secciones es mostrar los ajustes de los modelos aditivos y aditivos multiplicativos con las mismas variables con el fin de determinar cuáles son las que mejor se ajustan en cada caso y así poder determinar más alternativas.

Modelo de riegos aditivos

Dado que el ajuste de la variable *chla* provoca una ligera discrepancia de los datos con el modelo proporcional, se investiga sobre su ajuste con un modelo aditivo. Este modelo no tiene ningún supuesto parecido al de proporcionalidad modelo anterior, por lo que las pruebas que se harían en este caso no se realizarán con el objetivo de corroborar que se cumplan ninguna hipótesis, si no que se apunta hacia el objetivo de investigar el efecto invariante en el diagnóstico del paciente y a evaluar la congruencia del modelo.

Ajuste del modelo

Primeramente se ajusta un modelo aditivo no paramétrico con todas las variables incluidas en el modelo **3.1.2** anterior. En el **cuadro 3.10 a)** se puede observar que la interacción *less17:yschool* y la transformación *log(npartner)* no son significativas con un nivel de confianza del 95 %. Por ello se busca un conjunto de variables más pequeño para incorporar en el modelo. Finalmente, en el **cuadro 3.10 b)** se observan los resultados con las variables más significativas.

Cuadro 3.10: Prueba de significancia para las variables en el ajuste de un modelo aditivo con las variables *chla*, *os12m*, *condom*, *log(npartner)* y *less17*yschool* en el cuadro **a)**. Misma prueba en el cuadro **b)** omitiendo *log(npartner)* y la interacción con *less17*.

a)	$sup\|B(t)/\Phi(t)\|$	*P-valor*
intercepto	6.42	0.001
chla	3.45	0.027
os12m	5.68	8.2×10^{-4}
condom	3.25	0.023
log(npartner)	2.55	0.186
less17:yschool	2.75	0.116

b)	$sup\|B(t)/\Phi(t)\|$	*P-Valor*
intercepto	7.28	$< 10^{-4}$
chla	3.38	0.021
os12m	5.13	$< 10^{-4}$
condom	3.27	0.029
yschool	4.05	2.7×10^{-4}

Determinado el modelo del cuadro anterior, se analiza el siguiente ajuste:

$$\lambda(t) = \beta_0(t) + \beta_1(t) \cdot chla + \beta_2(t) \cdot os12m + \beta_3(t) \cdot condom + \beta_4(t) \cdot yschool \qquad (3.1.4)$$

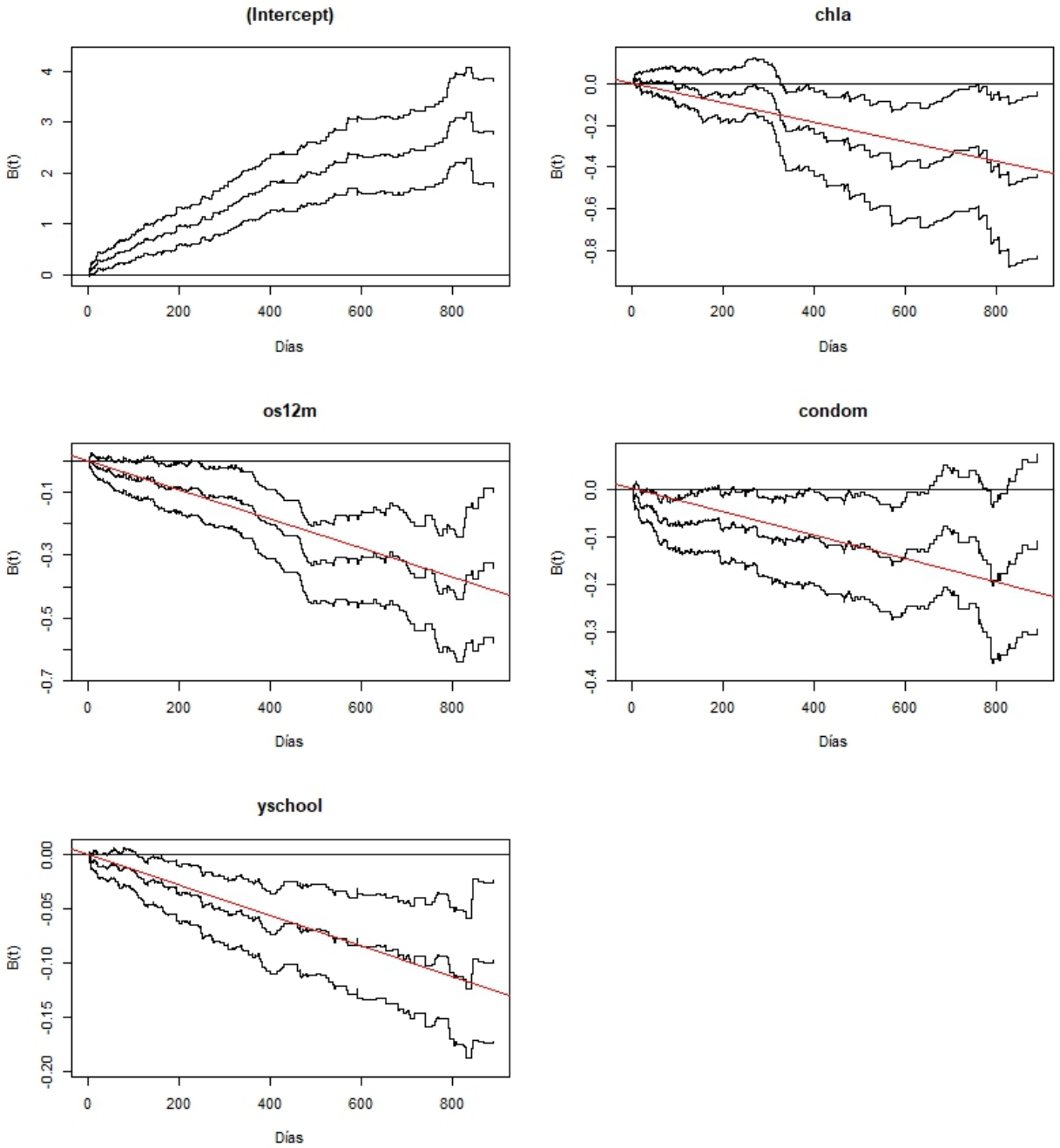
Se procede a hacer el análisis de Kolmogorov-Smirnov para determinar si existen variables con efectos variables en el diagnóstico. La función dela paquete *timereg*, *aalen()*, permite observar los resultados de esta prueba para contrastar:

$$H_0 : B_j(t) = \gamma t \quad \text{vs.} \quad H_1 : B_j(t) \neq \gamma t \qquad (3.1.5)$$

los cuales se muestran en el **cuadro 3.11**.

Por otro lado, en la **fig. 3.18** se muestran los coeficientes de regresión acumulados, los cuales no dan indicios de tener un efecto variable a lo largo del tiempo. El **cuadro 3.11** justamente reporta que ninguna variable parece tener un efecto significativamente variante (bajo un nivel de significancia del 10 %).

Figura 3.18: Coeficientes de regresión acumulados $B(t)$ para cada una de las variables del modelo **3.1.4**.

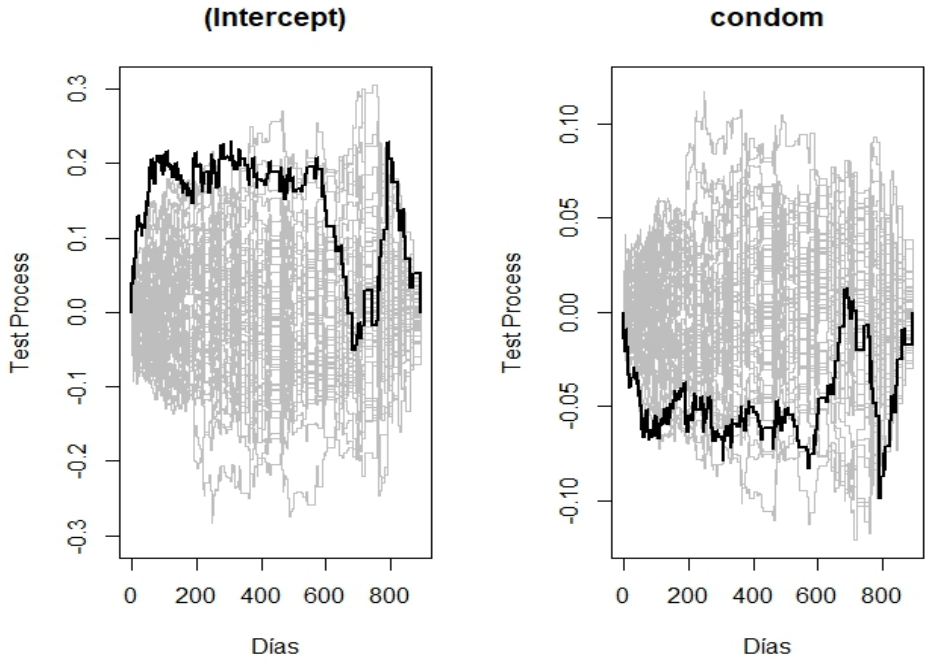


Cuadro 3.11: Prueba de Kolmogorov-Smirnov para contrastar las hipótesis **3.1.5** de efecto invariante.

	Kolmogorov	*P-valor*
intercept	0.6210	0.114
chla	0.1320	0.634
os12m	0.1450	0.154
condom	0.1040	0.152
yschool	0.0318	0.425

Puede observarse también los resultados del *Test process* para probar el efecto invariante de las variables, en este caso para *condom*, que resultó acercarse a ser de efecto variable según la prueba de Kolmogorov-Smirnov. Así, efectuando dicho proceso, se muestran en la **fig. 3.19** las curvas de las simulaciones bajo la hipótesis nula de **3.1.5** junto con la curva del modelo actual. Tal parecer ser que no se desvían significativamente las observaciones de las simulaciones bajo la hipótesis nula. Por lo tanto, se consideran las covariables de efecto constante. El intercepto $B_0(t)$ también parece no ser variante.

Figura 3.19: *Test process* para la prueba de efecto constante de la variable *condom* en el modelo **3.1.4** junto con 50 simulaciones bajo la hipótesis nula.



Por los resultados anteriores, se decide trabajar con un modelo aditivo completamente paramétrico. Es decir,

$$\lambda(t) = \beta_0(t) + \beta_1 \cdot chla + \beta_2 \cdot os12m + \beta_3 \cdot condom + \beta_4 \cdot yschool \tag{3.1.6}$$

Esto además, para poder dividir a las pacientes en grupos de riesgo. Una vez ajustado el modelo paramétrico, se tienen en el **cuadro 3.12** los coeficientes estimados del modelo.

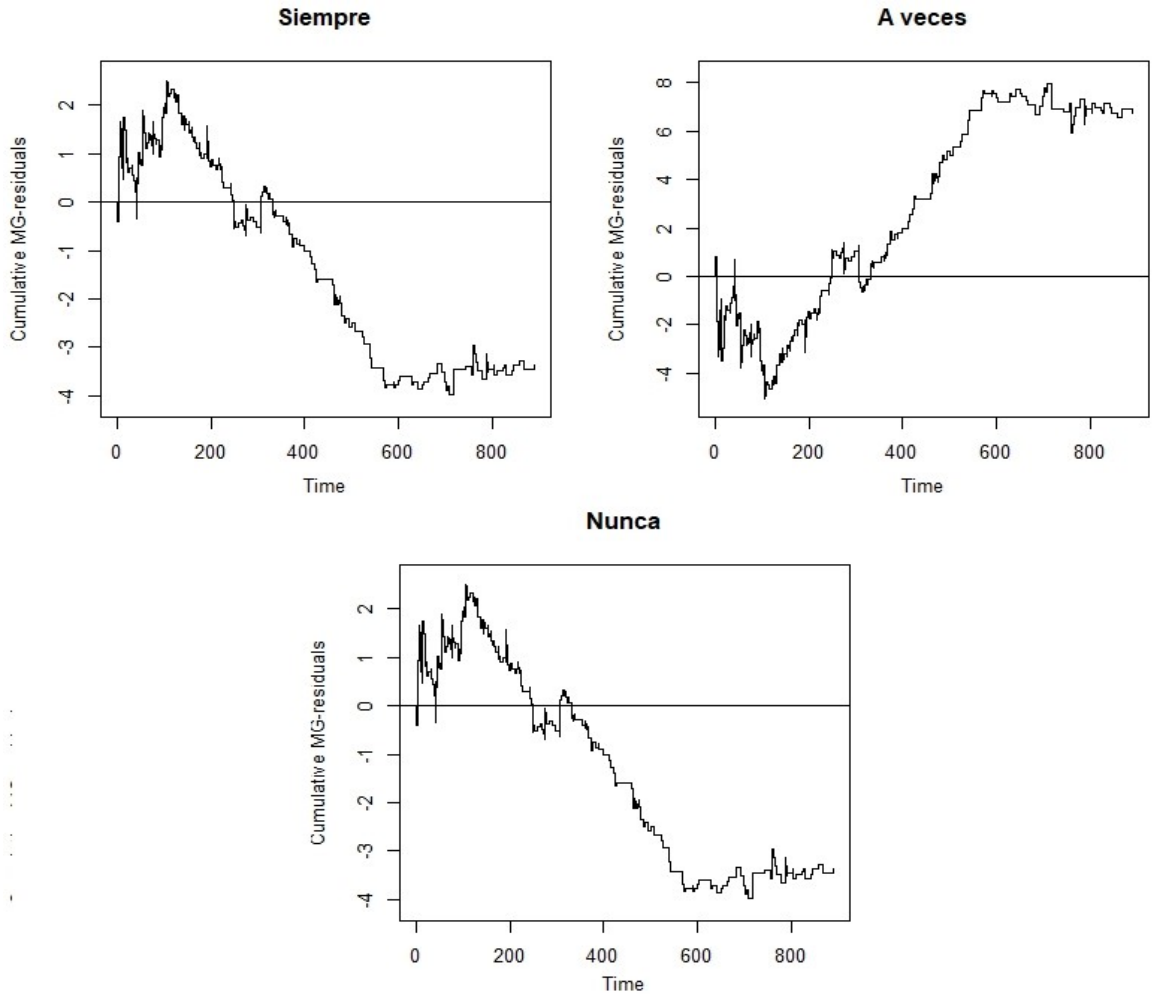
Procedimientos de bondad de ajuste

Cuadro 3.12: Coeficientes estimados del modelo aditivo paramétrico **3.1.4**.

	$\beta_i \times 10^4$	β_i/σ_i	*P-valor*
chla	-4.67	-2.41	0.016
os12m	-4.61	-3.83	$< 10^{-3}$
condom	-2.44	-2.26	0.024
yschool	-1.41	-3.80	$< 10^{-3}$

Nuevamente se estiman las martingalas residuales con el objetivo de observar su comportamiento para poder determinar si efectivamente se ajustan los datos a la predicción del modelo. La forma en la que se llevará acabo este procedimiento será, en esta ocasión para el modelo paramétrico, dividir cada variable en estratos determinados con los cuartiles para las variables continuas y niveles para las variables discretas. Posteriormente graficar los residuales de martingala acumulativos del modelo completo vs. el tiempo, ya que al estratificar se podrá ver el comportamiento, no sólo a lo largo del tiempo, si no también en cada cuantil o nivel en el ajuste. Los residuales se calculan para cada variable, aunque sólo se muestran para la variable *condom*, pues es aquella que parece tener las desviaciones más relevante.

Figura 3.20: Residuales de martingalas acumulativos para la covariable *condom* del modelo **3.1.6**.

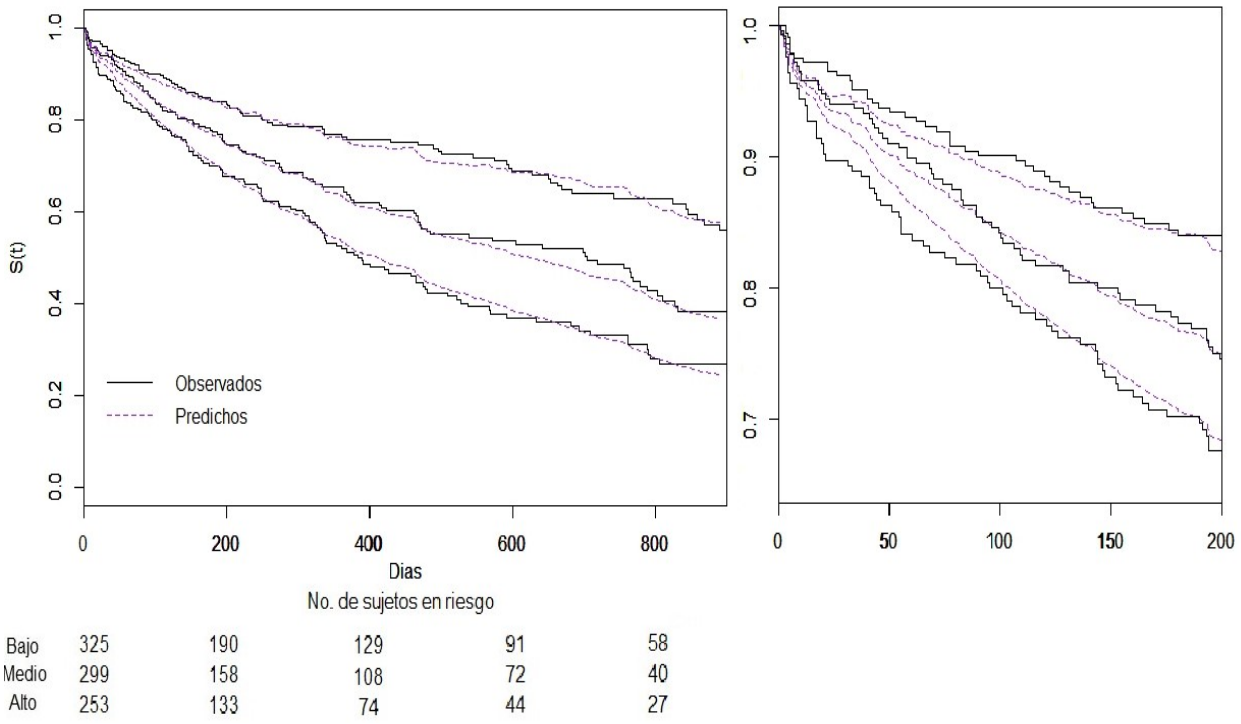


Lo que se puede observar de la **fig. 3.20** es que, al parecer se tiene bien descrito el modelo de riesgos aditivos mediante la variable *condom* durante la primera mitad del intervalo de estudio. Para los niveles correspondientes a las pacientes que siempre y nunca usaron condón, el modelo tendió a predecir menos muertes conforme pasaba el tiempo. Fue para las pacientes que a veces usaban condón, que cada vez se predecían más reinfecciones que las que realmente ocurrieron. En general se tiene una buena predicción por parte del modelo.

Grupos de riesgo

Gracias a que se cuenta con un modelo enteramente paramétrico, es posible realizar la clasificación de los grupos de riesgo para las pacientes de la población. En la **fig. 3.21** se muestran los estratos de las observaciones divididos por grupos de riesgo comparados con las predicciones del modelo dadas por la función estimada evaluada en las medias de cada variable restringida al grupo de riesgo correspondiente.

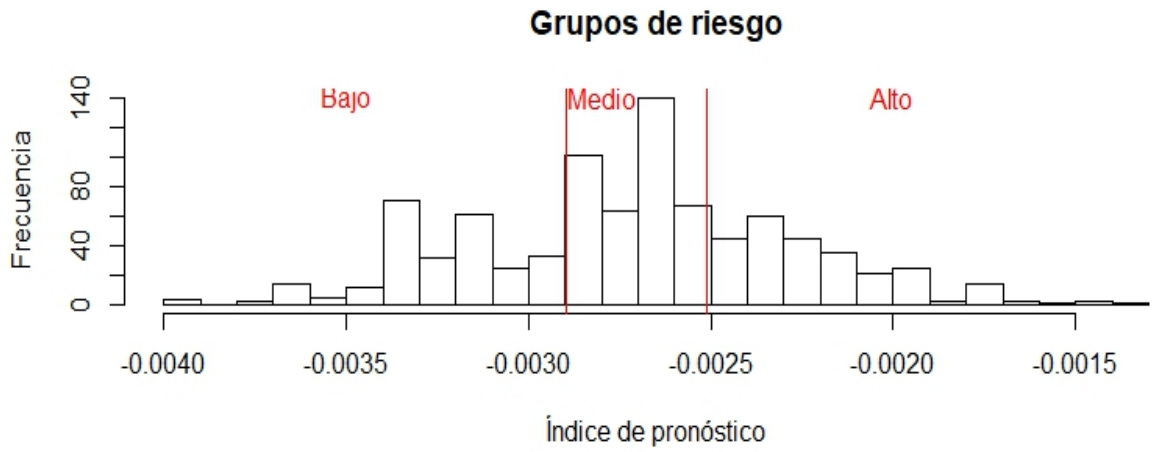
Figura 3.21: Estimador de Kaplan-Meier estratificado por grupos de riesgo, comparado con la función de supervivencia del modelo aditivo **3.1.6** evaluada en la media de cada variable restringida a cada grupo.



No. de sujetos en riesgo

Bajo	325	190	129	91	58
Medio	299	158	108	72	40
Alto	253	133	74	44	27

Figura 3.22: Histograma para el índice de pronóstico del modelo **3.1.6** con grupos de riesgos señalados por la líneas rojas en los terciles del índice.



Al respecto, se puede decir que se tiene una clasificación aproximadamente igual de buena que aquella dada por el modelo proporcional, ya que la prueba de log rank arroja un p-valor de $4.1\text{x}10^{-9}$ (valor cercano al obtenido en la clasificación anterior). En la **fig. 3.22** se observan los grupos de riesgos representados en el histograma del índice de pronóstico.

Ajuste de Modelo aditivo-multiplicativo

Luego de haber realizado los ajustes de los modelos anteriores, se pudo ver que en cada modelo habían variables que no se ajustaban correctamente a los datos acorde al modelo. Para el ajuste del modelo aditivo-multiplicativo, se considerará explorar en las transformaciones aplicadas a las variables, así como la correcta identificación de las variables no proporcionales dentro del modelo.

Dado que en el modelo de riesgos proporcionales, la variable *chla* resultó no dar evidencia de ser proporcional en el modelo (ver **fig. 3.12**), se considera en esta ocasión asignarle un lugar en la parte aditiva del modelo, la cual tendrá un efecto variable en el riesgo de las pacientes. Por otro lado, se piensa comparar la transformación de la variable *npartner*. Con respecto a los ajustes, en la **fig. 3.23** se muestran los coeficientes de regresión acumulados $\hat{B}(t)$ para el intercepto y la variable *chla* de los siguientes dos modelos.

$$\lambda(t) = \{\beta_0(t)+\beta_1(t)\cdot chla\}\cdot exp(\gamma_0\cdot os12m+\gamma_1\cdot condom+\gamma_2\cdot npartner+\gamma_3\cdot yschool\cdot less17) \tag{3.1.7}$$

$$\begin{aligned}\lambda(t) = \{\beta_0(t) + \beta_1(t)\cdot chla\}\cdot exp(\gamma_0\cdot os12m + \gamma_1\cdot condom+ \\ \gamma_2\cdot \log(npartner) + \gamma_3\cdot yschool\cdot less17)\end{aligned} \tag{3.1.8}$$

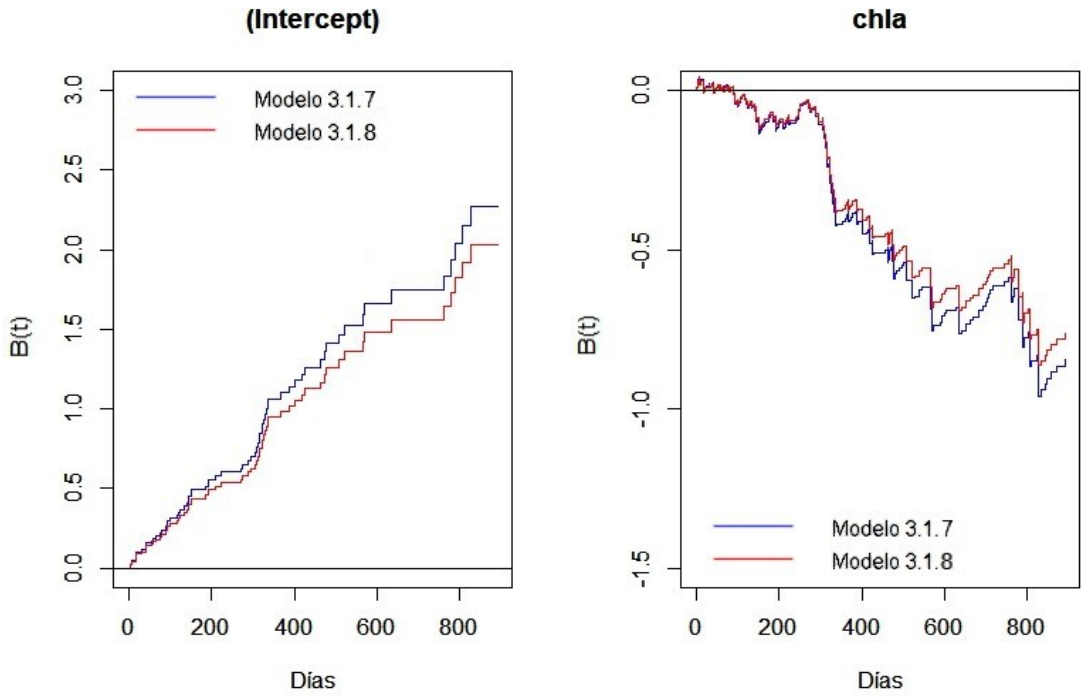
Cuadro 3.13: Prueba de significancia para la parte proporcional del modelo con *log(npartner)* (izquierda) y *npartner* (derecha) de los modelos **3.1.7** y **3.1.8**

a)	$\hat{\beta}_i$	$\hat{\beta}_i/\sigma_i$	*P-valor*	**b)**	$\hat{\beta}_i$	$\hat{\beta}_i/\sigma_i$	*P-valor*
os12m	-0.603	-3.76	$< 10^{-3}$	*os12m*	-0.571	-4.16	$< 10^{-3}$
condom	-0.207	-2.11	0.035	*condom*	-0.195	-1.97	0.049
nparnter	0.097	1.83	0.068	log(*npartner*)	0.286	1.97	0.048
yschool:less17	0.040	2.84	0.004	*yschool:less17*	0.039	2.82	0.005

Las diferencias en estos modelos, al considerarse la transformación $\log(npartner)$, pueden contribuir a tener un modelo que mejor describa los datos. Obsérvese que se tiene una ligeramente mayor significancia cuando se tiene presente la transformación *log(nparter)* en el **cuadro 3.13** (menor a 5 %). Por ello es importante evaluar la bondad de ajuste del modelo.

Por otro lado, es interesante observar que la variable *chla* es la que tiene efecto aditivo en el modelo. Y como se había mencionado en el **capítulo 2**, la inclusión de esta variable funciona como una estratificación en el modelo paramétrico de riegos proporcionales. Es decir, para un individuo con *condom*, log(*npartner*) y *less17* iguales a cero, se tiene una función de riesgo de base que depende del valor de *chla* y del tiempo para cada individuo. En la **fig. 3.23** se muestran los coeficientes del vector $B(t)$ para el intercepto y *chla* del modelo **3.1.7** y **3.1.8**.

Figura 3.23: Coeficientes de regresión acumulados para el intercepto y la variable *chla* del modelo **3.1.7** (rojo) y del modelo **3.1.8** (azul).



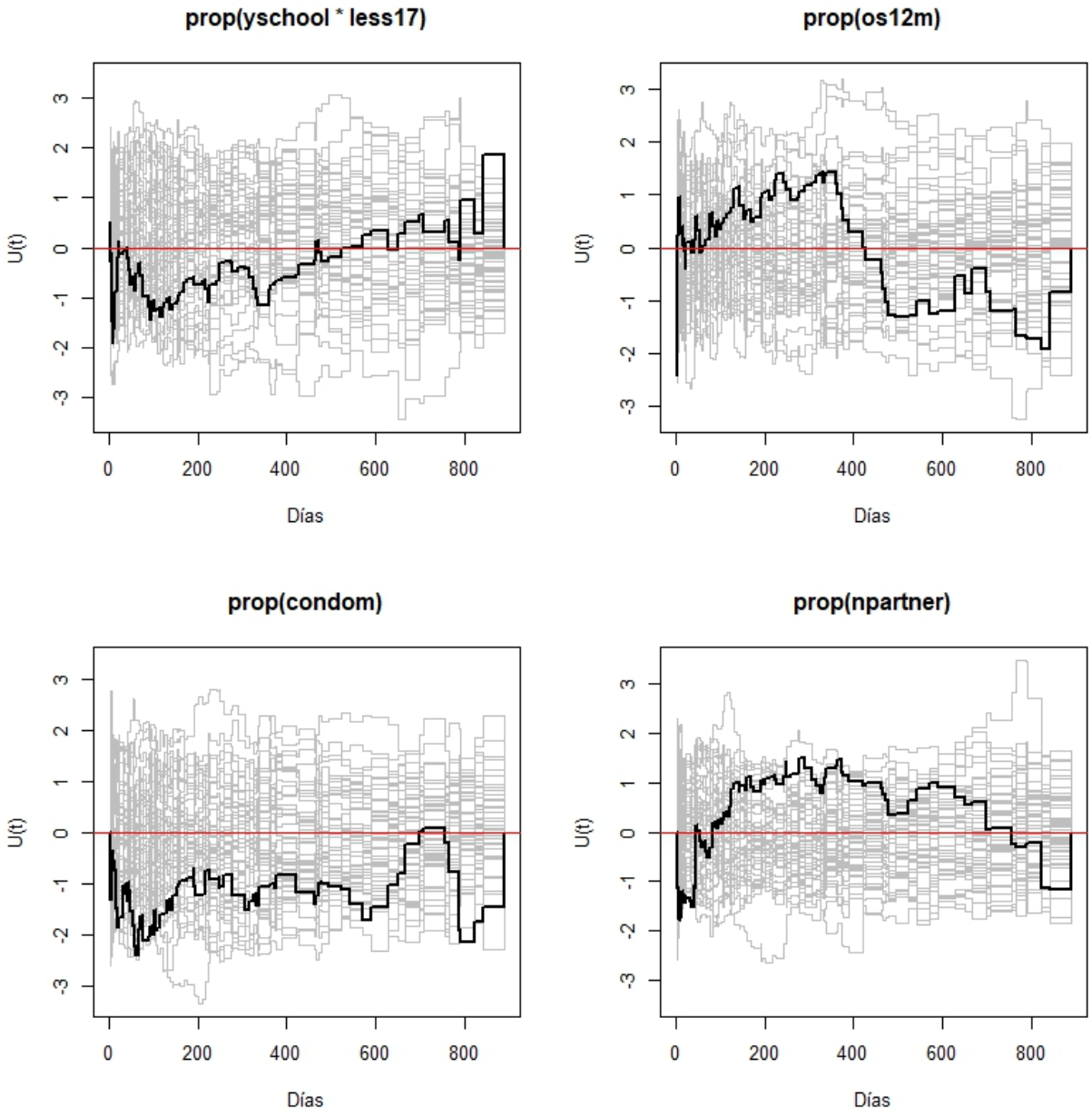
Cuadro 3.14: Prueba de proporcionalidad del modelo **3.1.8** según el *score process*.

	$sup\|U(t)\|$	*P-valor*
os12m	2.41	0.222
condom	2.43	0.220
log(*npartner*)	2.64	0.162
less17:yschool	1.94	0.554

Prueba de proporcionalidad

Esta parte del análisis se muestra para el modelo **3.1.7**, ya que el modelo **3.1.8** presenta resultados idénticos. De tal modo, para corroborar que las variables proporcionales del modelo sí cumplen el supuesto de proporcionalidad, se realiza el *Score process* para comparar las simulaciones bajo la hipótesis nula con la curva de la función score. Adicionalmente, en el **cuadro 3.14** se tiene la prueba formal para contrastar la hipótesis de proporcionalidad y la **fig. 3.24** sirve de apoyo al respecto. Se concluye que ninguna variable carece de efecto proporcional.

Figura 3.24: Proceso de score para las variables proporcionales del modelo **3.1.7** junto con 50 simulaciones bajo el supuesto de proporcionalidad.



Prueba de Bondad de Ajuste

Se estiman los residuales de martingalas acumulativos de los modelos **3.1.7** y **3.1.8** y se analiza, para cada variable, los estratos definidos por sus cuantiles. Se fija la atención especialmente para la variable *npartner* pues se alterna la transformación *log()* en los modelos **3.1.7** y **3.1.8**. Los niveles con los que se analiza la bondad de ajuste son simplemente definidos por los valores $\{0, 1\}$ que corresponde a los casos de escasas parejas sexuales y $(1, 19]$ para los casos con varias parejas sexuales. Luego se compararán los residuales obtenidos para cada cuantil con 50 realizaciones simuladas de las martingalas residuales bajo el modelo supuesto.

Figura 3.25: Residuales de martingalas acumulativos observados para los niveles de *npartner* del modelo **3.1.7** (0 y 1 parejas componen 'escasos' y 2 o más son 'varios').

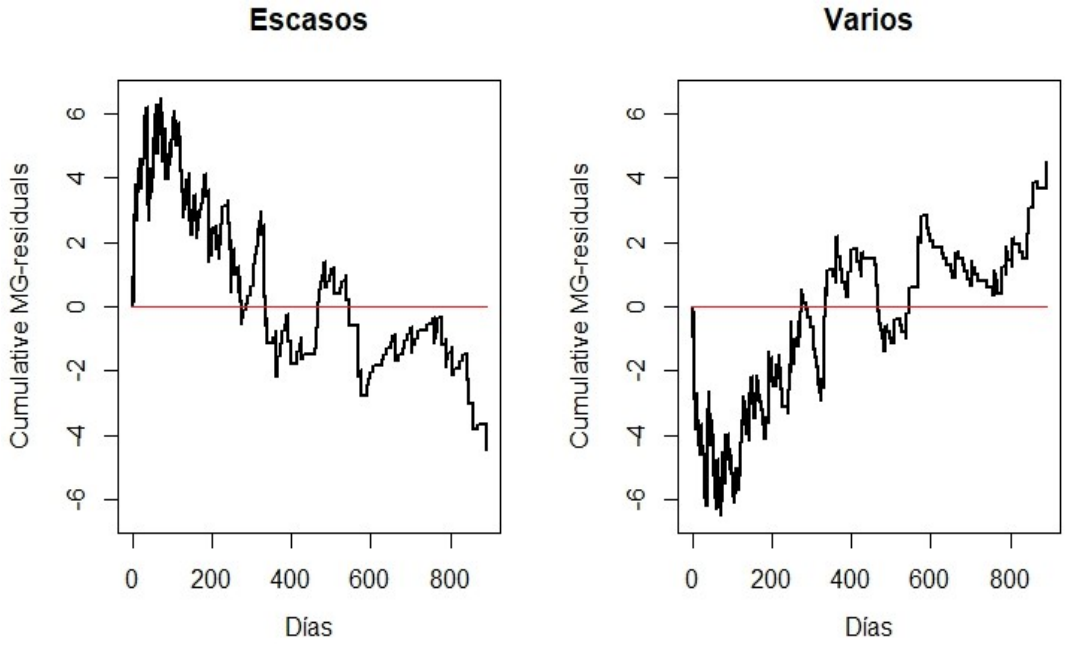
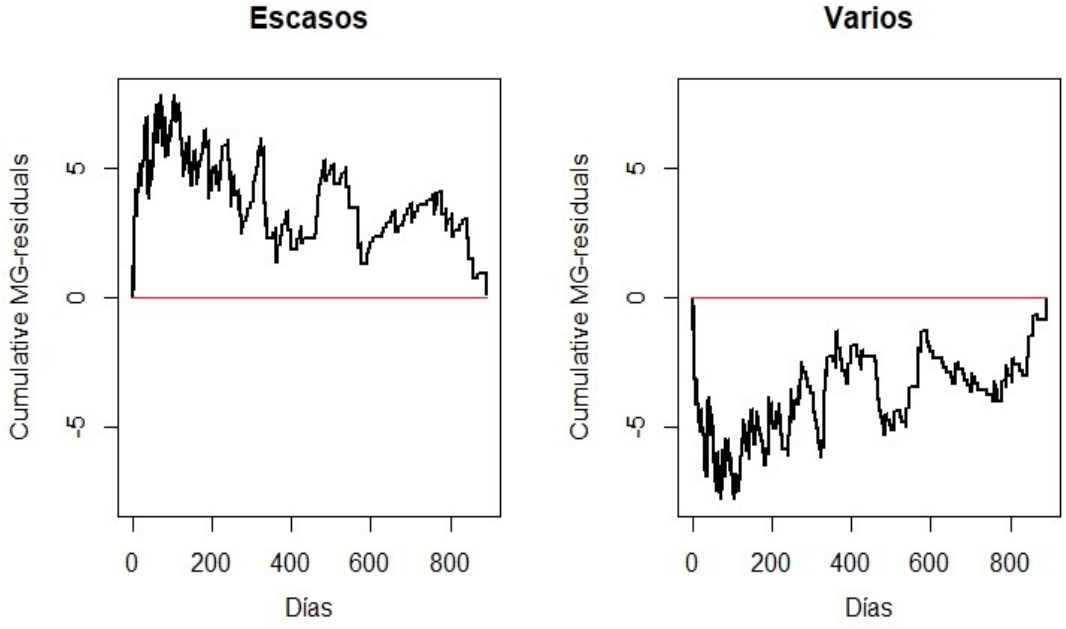


Figura 3.26: Residuales de martingalas acumulativos observados para los niveles de log(*npartner*) del modelo **3.1.8**.

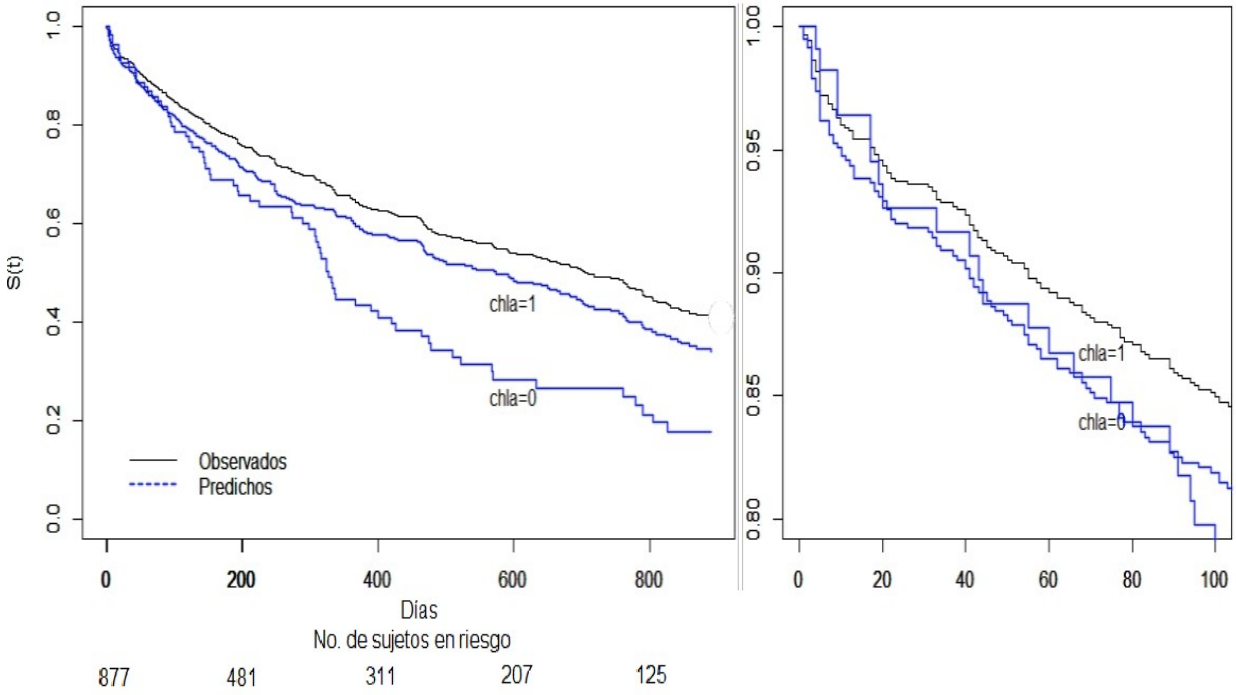


Se puede concluir que realmente se tiene una capacidad predictiva ligeramente más adecuada para el modelo **3.1.7** sin la transformación se realiza en *npartner*, dado que los residuales indican pequeñas desviaciones de las predicciones.

Estimación de la función de supervivencia

En este caso, sólo la variable *chla* hace este modelo diferente al modelo de riesgos proporcionales. A pesar de ello, no es recomendable crear un índice de riesgo con base en las covariables proporcionales. Por ello, esta sección se limitará a mostrar la función de supervivencia estimada, estratificada según la variable *chla* y evaluada en las medias de las variables *os12m*, *condom*, *partner* y la interacción *less17:yschool*. Además, se comparan dichas curvas con la estimación de Kaplan-Meier, a modo de observar qué tanto se aproximan las estimaciones del modelo al comportamiento de la supervivencia de la población según los datos observados.

Figura 3.27: Función de supervivencia estimada estratificada por la variable *chla* y evaluada en las medias de las variables *os12m*, *condom*, *partner* y la interacción *less17:yschool* (valores predichos). Estimador de la función de supervivencia de Kaplan-Meier (valores observados).



El modelo parece predecir un mayor riesgo sin importar la presencia de clamidia en una paciente con características estándares.

Conclusiones

Se puede observar que, en cuanto a los modelos proporcionales o aditivos, parecen distinguir adecuadamente los grupos de riesgo bajo, mediano y alto. Si bien dichos modelos pueden no ajustarse del todo adecuadamente a las observaciones en todos los casos, se pueden clasificar los grupos de riesgo. Parecería que algunos resultados son un poco distintos a lo que se esperaría del modelo, pero estos dependen de la representatividad de los datos que, en todo caso, sólo describen el conjunto de datos de la base *std*. Sin embargo, con respecto a los modelos ajustados se puede decir lo siguiente:

- El modelo de riesgos proporcionales parece haber brindado una estimación de la función de supervivencia más cercana a la estimación de Kaplan-Meier con las observaciones. Por lo tanto, podría decirse que es el modelo más adecuado para ajustar los datos.
- A pesar de que el modelo de riesgos aditivos podría no parecer más preciso, la interacción de la edad con respecto a los años de estudio se comporta de forma interesante. Se puede ver, en el **cuadro 3.4 b)** correspondiente al modelo proporcional, que la interacción de las covariables *less17* y *yschool* parece propiciar un mayor riesgo de reinfección en los pacientes. Sin embargo en las estimaciones del **cuadro 3.12** correspondiente al modelo aditivo, se tiene el efecto opuesto al no considerar la interacción con la edad y sólo incorporar al modelo la covariable *yschool*. Una posible explicación a esto es que en edades menores a 17 años, en los años de estudio se observa generalmente una exposición a un mayor número de parejas sexuales potenciales, dado que en la educación media superior se presenta esa posibilidad. Sin embargo, al no considerar la interacción con la edad, el haber estudiado un tiempo considerable, favorece la concientización sobre el riesgo de contagio de enfermedades de transmisión sexual.
- A pesar de que el modelo de riesgos aditivo-multiplicativo no posee una capacidad predictiva tan precisa como los modelos de riesgos proporcionales o aditivos, reafirma el contraste que existe entre las pacientes que padecieron clamidia y quienes no. Esto es debido a que la covariable binaria *chla* se incorporó al modelo en la parte aditiva.

Para identificar los grupos de la población más vulnerables a una reinfección, se muestran en los cuadros **3.15** y **3.16** las medias de cada variable restringida a cada grupo de riesgo determinado por los terciles del índice de pronóstico para los modelos de riesgos proporcionales y aditivos. Justamente en estos valores se evaluaron las estimaciones de la función de supervivencia para los modelos de riesgos proporcionales y aditivos, cuyas funciones de riesgo se muestran a continuación respectivamente:

$$\begin{aligned}\lambda(t) = \lambda_0(t) \cdot exp(\beta_1 \cdot chla + \beta_2 \cdot os12m + \beta_3 \cdot condom + \\ \beta_4 \cdot \log(npartner) + \beta_5 \cdot less17 \cdot yschool)\end{aligned} \tag{3.1.9}$$

$$\lambda(t) = \beta_0(t) + \beta_1 chla + \beta_2 os12m + \beta_3 condom + \beta_4 yschool$$

Cuadro 3.15: Valores medios de cada variable correspondientes a cada grupo de riesgo para el modelo de riesgos proporcionales de la expresión **3.1.2** en los que se evaluó la función de supervivencia estimada de la **fig. 3.16**. A la covariable *npartner* se le aplicó la transformación *log()* en el ajuste, por lo que se muestra la transformación inversa de la media evaluada para su interpretación.

	chla	*os12m*	*condom*	*npartner*	*less17:yschool*
Bajo	0.958	0.657	2.6	1.144	0.449
Medio	0.926	0.117	2.04	1.12	0.151
Alto	0.621	0.038	2.062	1.39	4.166

Es importante observar que el promedio de la interacción *less17:yschool* se interpreta como el promedio de años de estudio de las pacientes que son menores de 17 años.
Conforme a los resultados de la variable *npartner* correspondientes al número de parejas, se observa en el cuadro anterior que en promedio, las personas con alto riesgo de contraer una reinfección sólo tienen entre una y dos parejas sexuales. Lo cuál da una impresión de qué tan fuerte es la presencia de la enfermedad en la población. Con ello, la opinión médica puede emitir juicios sobre esos resultados.

Cuadro 3.16: Valores medios de cada variable correspondientes a cada grupo de riesgo para el modelo aditivo de la expresión **3.1.6** en los que se evaluó la función de supervivencia estimada de la **fig. 3.21**.

	chla	*os12m*	*condom*	*yschool*
Bajo	0.95	0.652	2.56	12.596
Medio	0.899	0.227	2.197	11.351
Alto	0.557	0.031	2.067	10.007

En el cuadro anterior, se puede interpretar que para esta población las personas más vulnerables son aquella que no tuvieron sexo oral a largo plazo, si no a corto plazo. Esto se puede ver en la variable *os12m* que señala los pacientes que tuvieron sexo en los últimos meses.

Finalmente, se muestra en el **cuadro 3.17** las medias con las que se evaluó la función de supervivencia estimada del modelo de riesgos aditivo-multiplicativo acorde a la siguiente función de riesgos :

$$\lambda(t) = \{\beta_0(t) + \beta_1(t)chla\} \cdot exp(\gamma_0 \cdot os12m + \gamma_1 \cdot condom + \gamma_2 \cdot npartner + \gamma_3 less17 \cdot yschool) \tag{3.1.10}$$

Cuadro 3.17: Valores medios de todas las observaciones para cada covariable de la parte proporcional del modelo multiplicativo-aditivo de la expresión **3.1.7**. Corresponde a los puntos de evaluación de la función de supervivencia estimada de la **fig. 3.27**.

os12m	*condom*	*npartner*	*less17:yschool*
0.328	2.294	1.346	1.6

Estos resultados permitirán enriquecer el diagnóstico médico de esta población de pacientes y mostrar evidencia para los grupos poblacionales más vulnerables. De tal modo que se pueda esclarecer cuáles serían las acciones más convenientes para poder prevenir una reinfección de gonorrea y clamidia, las cuales son enfermedades que es preciso erradicar.

3.2. Estudio de cirrosis biliar primaria

Modelo de riesgos proporcionales paramétrico

Como segundo ejemplo, se procede a hacer el análisis de la base de datos *pbc* (Primary Billiar Cirrosis), que contiene los datos de 424 pacientes a los cuales se les dio seguimiento de Enero de 1974 hasta Mayo de 1984. A 312 pacientes se les administró aleatoriamente una sustancia activa llamada D-penicilamina (DPCA). A estos mismos pacientes se les recolectó información de variables clínicas, bioquímicas, serológicas e histológicas. Del total de 312, 125 murieron a la fecha de fin de estudio con sólo 11 muertes no atribuibles al PBC. Una operación de transplante de riñón se hizo para 19 pacientes de este grupo. El otro grupo de 112 pacientes restantes no recibió la administración de DPCA y de ellos, a 6 se les perdió el seguimiento apenas inició el estudio, 36 murieron para junio de 1986 y 6 se sometieron a un trasplante de hígado. La separación de los 312 pacientes se llevó a cabo para probar la efectividad de la DPCA en el tratamiento contra el PBC.

Descripción de los datos

Las variables que conforman la base de datos se clasifican del siguiente modo: *demográficas* para describir el grupo poblacional al que pertenece cada individuo, *clínicas* para describir algunas características observadas en el diagnóstico inicial del paciente durante su internación, *bioquímicas* para referirse al nivel de los componentes que el médico determinó después de la examinación e *histológicas* para las condiciones en las que se encuentra el tejido del riñón del paciente.

Cuadro 3.18: Descripción de variables incluidas en la base de datos *std* del paquete *KMsurv* en R.

Nombre	Descripción	Codificación/Unidades
days	Días desde que se registra el paciente hasta su muerte, se le realiza un transplante de riñón o llega al fin del estudio en julio de 1986 (estos últimos dos casos se consideran censura)	Días
status	Indica si se censuró a observación. Un astedisco si se le hizo transplante.	1: si ocurrió una muerte por PBC. 0: en otro caso.
Variables demográficas		
trt	El tratamiento asignado.	1: para D-penicilamina. 2: para placebo.
age	Edad del paciente	Años
sex	Sexo del individuo	1: si es mujer. 0: si es hombre.

Variables clínicas		
ascites	Presencia de ascitis (fluido peritoneal)	1: presente. 0: ausente.
hepatomegaly	Hepatomegalia (crecimiento anormal del hígado)	1: presente. 0: ausente.
spiders	Várices y otras malformaciones sanguínea.	1: presente. 0: ausente.
edema	Hinchazón por un exceso de líquido en algún tejido. Indica su presencia en relación a una terapia diurética corrientemente administrada.	0: Ausente. 0.5: Sin fallas en el tratamiento o sin tratar. Presente 1: Presente a pesar de la terapia.
Variables bioquímicas		
bili	Bilirubina sérica presente en la sangre.	mg/dl (miligramos por decilitros)
chol	Colesterol.	mg/dl (miligramos por decilitros)
albumin	Albúmina sérica presente en la sangre.	gm/dl (gramos por decilitros)
copper	Cobre presente en la orina.	μg/dl (microgramos por decilitro)
prothrombin	Tiempo en el que se produce la protrombina, proteína necesaria para coagular la sangre.	Segundos
platelet	Milésima parte del número de plataletas en 1 mm^3 de la sangre	Unidades
alk	Fosfatasa alcalina en la sangre.	U/ml (unidad enzimática por mililitro)
sgot	Aspartato aminotransferasa presente en la sangre.	U/ml (unidad enzimática por mililitro)
trig	Triglicéridos.	mg/dl (miligramos por decilitros)
stage	Estado de enfermedad del tejido hepático (1,2,3 o 4).	1: Primario. 2: Secundario. 3: Latente. 4: Final.

Analisis exploratorio de las variables

Para los 312 pacientes que fueron tratados con DPCA se tienen todos los datos asentados en las 19 variables del registro (salvo *cooper* y *platalet* que contienen algunos valores perdidos), mientras que los restantes 112 pacientes sólo consintieron el registro de algunas variables. En el **cuadro 3.18** se resume en la información de las variables que contienen todos los datos para los 312 individuos. Por otro lado, se muestran la distribución de las variables discretas presentes en el mismo cuadro para observar su distribución en la **fig 3.28**. De ellas se observa que la gran parte de la población es femenina. Algo parecido sucede con la presencia de ascitis. También se puede decir que la mayoría de los pacientes no presentaron edema pero sí tuvieron un avanzado grado de enfermedad en el tejido hepático. Aproximadamente a la misma cantidad de personas se les administró DPCA y Placebo.

Figura 3.28: Variables discretas de la base *pbc* y su distribución para sexo **a)**, ascitis **b)**, hepatomegalia **c)**, várices **d)**, edema **e)** y estado de enfermedad **f)**.

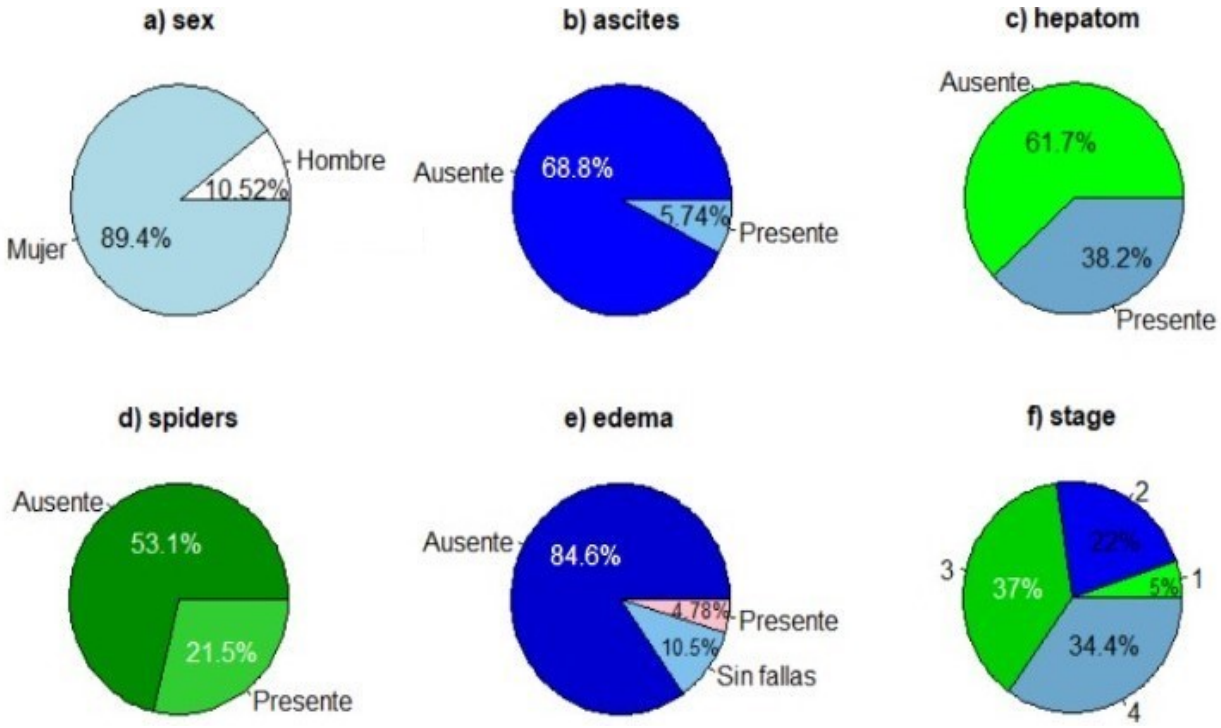
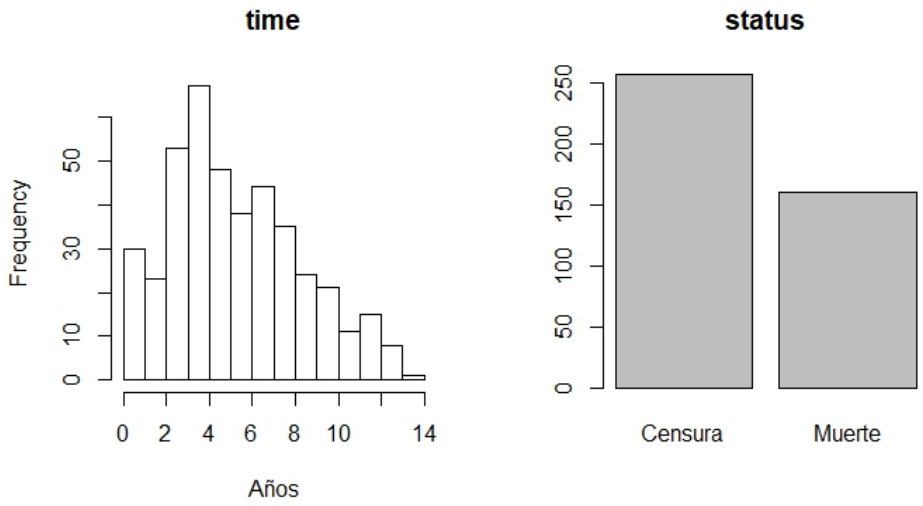
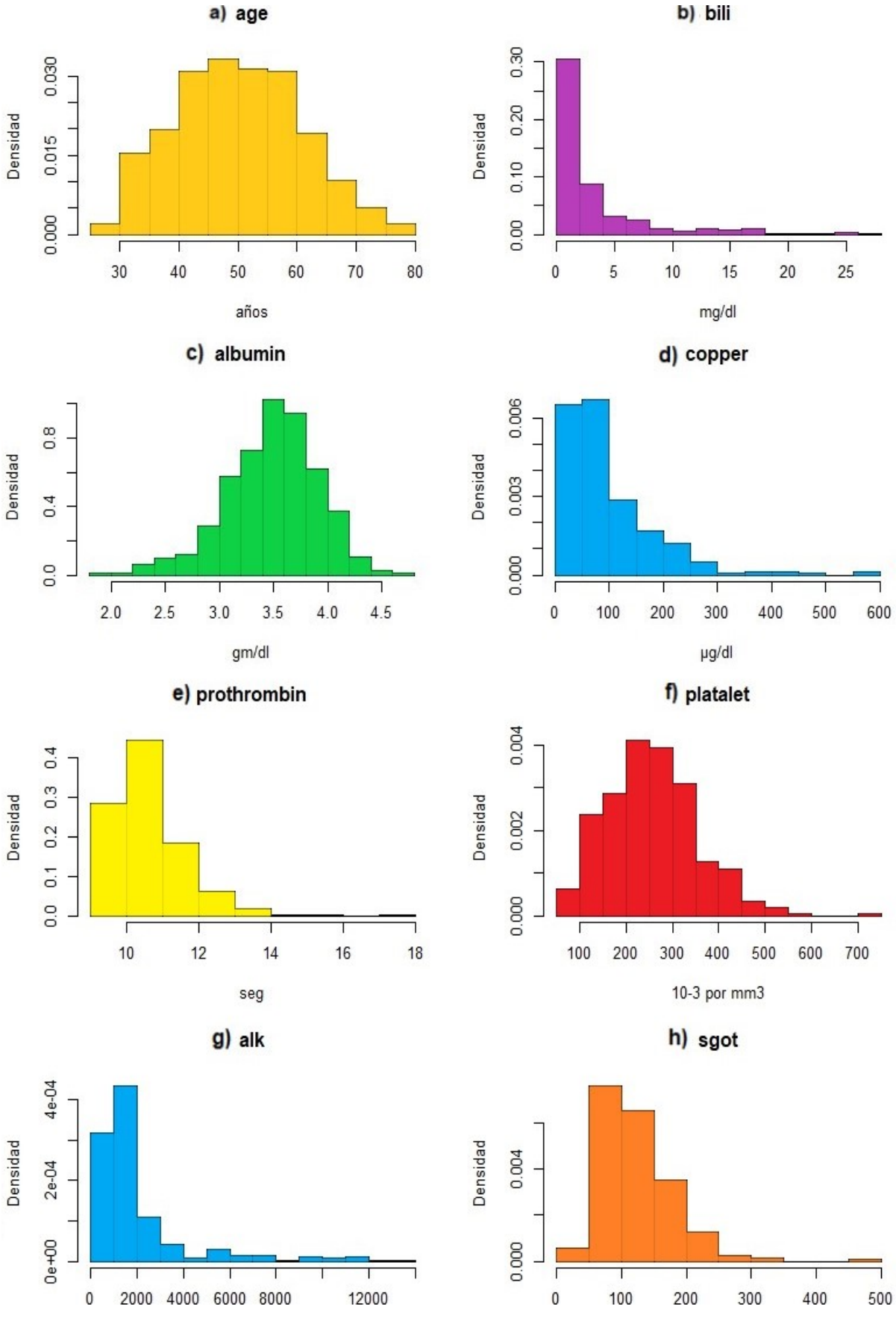


Figura 3.29: Distribución de la variable del tiempo *time* (izquierda) y censura *status* (derecha) de la base de datos *pbc*.



Acerca de las variables continuas, descritas en la figura **fig 3.30**, hay que mencionar que la edad promedio de los pacientes es alrededor de 50 años, la albúmina en la sangre está en un promedio de los 3.5 gm/dl y el resto de las variables tiene una gran concentración de sus datos en pequeñas cantidades. Algunos datos faltantes, como en la variable platalets, se sustituyeron por un promedio de los demás valores. La distribución de las variables del tiempo y censura se muestran en la **fig. 3.29**.

Figura 3.30: Variables continua de la base *pbc* y su distribución para edad **a)**, bilirubina **b)**, albumina **c)**, cobre **d)**, tiempo de coagulación **e)**, plataletas **f)**, alcalina **g)** y el estado de enfermedad **h)**.



Cuadro 3.19: Resumen de las 14 covariables para los datos de los 312 individuos con tratamiento aleatorio de DPCA de la base *pbc*.

	Variables	*Min*	*1. C*	*Med*	*3. C*	*Max*	*Datos faltantes*
Continuas	*age*	26.29	42.18	49.78	56.68	78.43	0
	bili	0.3	0.8	1.35	3.45	28	0
	albumin	1.96	3.31	3.55	3.8	4.64	0
	copper	4	41	73	123	588	2
	prothrombin	9	19	10.6	11.1	17.1	0
	platalet	62	200	257	323	563	4
	alk	289	867	1259	1985	13862	0
	sgot	28	81	115	152	457	0
	Variables	*Ausentes*		*Presentes*			*Datos faltantes*
Categóricas	*sex*	H: 36		M: 276			0
	ascites	288		24			0
	hepatom	152		160			0
	spiders	222		90			0
	trt	DPCA: 154		Placebo: 158			0
	edema	0: 263	0.5: 29	1: 29			0
	stage	1: 16	2: 67	3: 120	4:109		0

Dado que hay 312 individuos a los cuales se les administró aleatoriamente DPCA o Placebo, tiene sentido preguntarse si existe una diferencia entre los tiempo de falla de estos dos tipos de poblaciones, representados por las funciones de supervivencia $S_1(t)$ y $S_2(t)$ respectivamente. Así que con el fin de probar si existe una diferencia significativa entre ambos estratos poblacionales, se hace la estimación para la supervivencia de Kaplan-Meier estratificada entre los dos tipos de tratamiento. La estratificación se hace en la **fig. 3.31**.

Como se puede apreciar, no existe un efecto significativo para la población con DPCA sobre el resto. Es decir, el riesgo parece ser casi el mismo en estos 312 pacientes. Incluso, al realizar la prueba de Log Rank para contrastar la siguiente hipótesis

$$H_0 : S_1(t) = S_2(t) \text{ vs. } H_1 : S_1(t) \neq S_2(t)$$

la prueba arroja un p-valor de 0.75, por lo cual no se tiene suficiente evidencia para rechazar la hipótesis de que ambos estratos tienen la misma supervivencia. Esto sugiere que el grupo de 312 pacientes puede ser combinado para analizar posibles asociaciones con el tiempo de supervivencia y las medidas clínicas reportadas en la base *pbc*.

Figura 3.31: Estimador de Kaplan-Meier estratificado para el grupo que recibió DPCA y para el grupo control de la base *pbc*. Adicionalmente, se muestran las medianas y el número de pacientes que no habían presentado la falla a lo largo del tiempo para cada grupo y un aumento de la gráfica para los primeros años de estudio.

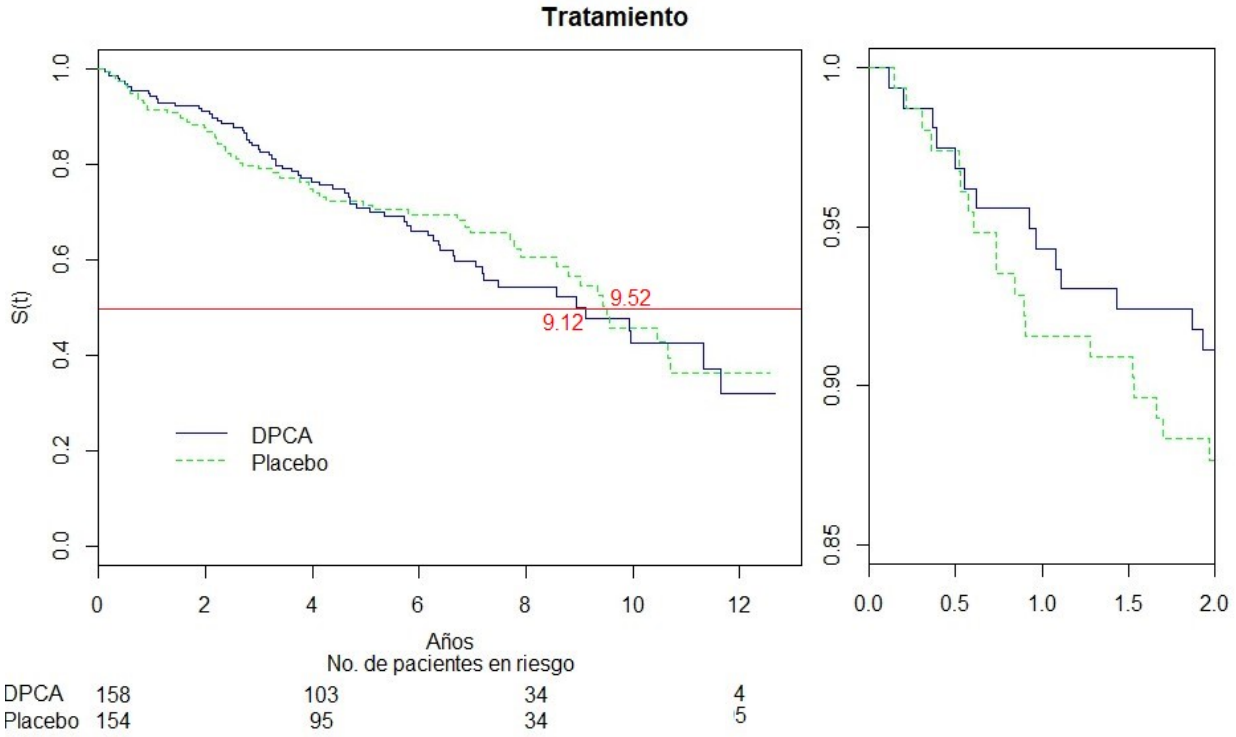
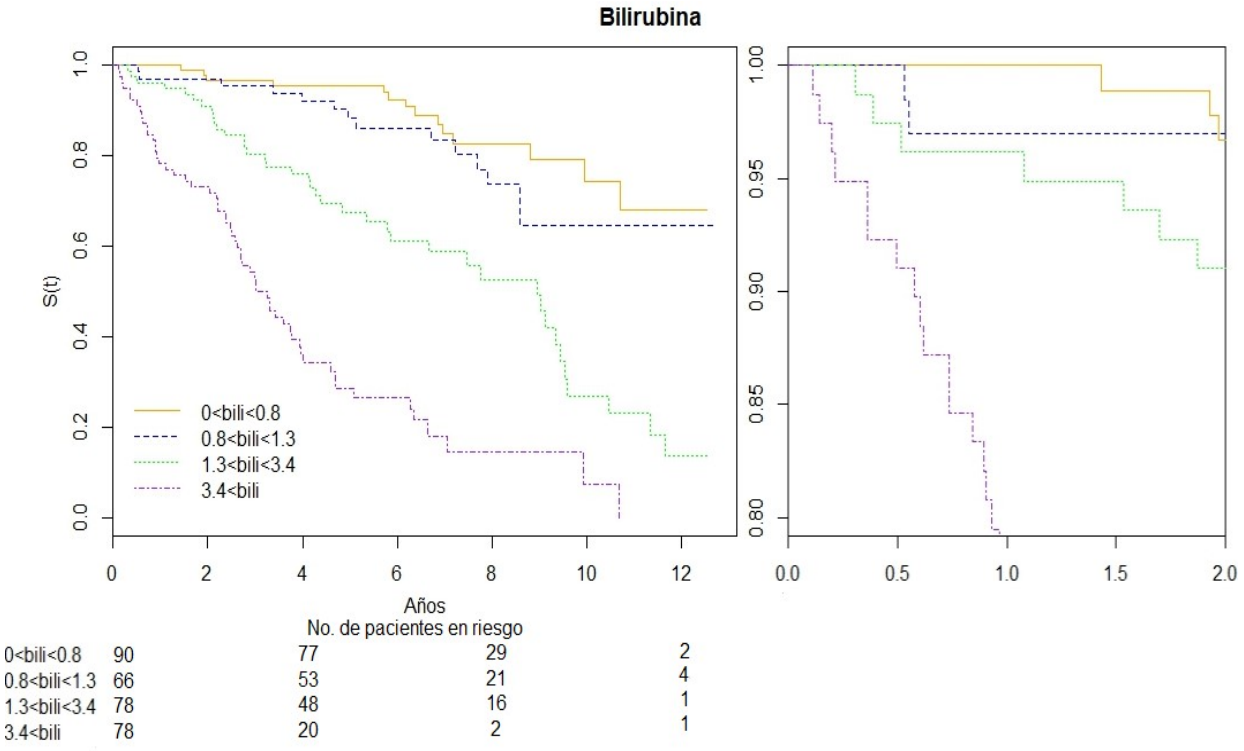


Figura 3.32: Estimador de Kaplan-Meier estratificado según los cuartiles de la variable *bili* en la base *pbc*. Se anexa el número de pacientes sobrevivientes en cada tiempo del intervalo de observación.



También es destacable la forma en la que se puede estratificar la población según el nivel de bilirubina que hay en la sangre, pues es la que mejor parece clasificar a los pacientes según su diagnóstico. Esto se puede apreciar en la **fig. 3.32**, donde la estratificación del estimador de Kaplan-Meier por los niveles definidos permite clasificar bastante bien los individuos acorde a su nivel de bilirubina. Esta variable también es la más significativa en el modelo como se verá en la siguiente sección.

Selección del modelo

Se hace el ajuste de un modelo de riesgos proporcionales paramétrico con el fin de describir el riesgo que existe para el grupo de pacientes con PBC. El primer modelo en ajustarse incluye 11 covariables de tipo demográficas, clínicas, bioquímicas e histológicas que se enlistan en el **cuadro 3.20 a)**. Posteriormente se usa la función *stepAIC()* para retirar variables que no son significativas en el modelo. La función arrojó los resultados mostrados en el **Cuadro 3.20 a) y b)** para el primer y último paso respectivamente.

Cuadro 3.20: Coeficientes estimados del modelo de riesgos proporcionales ajustado con las variables correspondientes en cada paso del algoritmo de selección de la función *stepAIC()* con los datos de la base *pbc* de los 312 pacientes a los que se les administró aleatoriamente DPCA.

a) *log verosim:*-550.3224	*Coef.* $\hat{\beta}_i \times 10$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
1. age	0.28	2.96	0.0026
2. albumin	-9.71	-3.62	0.0003
3. alk	1.4×10^{-4}	0.41	0.7023
4. ascites	2.82	0.91	0.3615
5. bili	1.06	6.34	$< 10^{-4}$
6. edema	6.94	2.14	0.0312
7. hepatom	4.88	2.22	0.0259
8. platelet	-6.06×10^{-3}	-0.59	0.5507
9. prothrombin	2.43	2.88	0.0041
10. sex	-4.76	-1.78	0.0737
11. spiders	2.89	1.38	0.1664
b) *log verosim:* -553.9490	*Coef.* $\hat{\beta}_i$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
1. age	0.034	3.69	0.0002
2. albumin	-1.072	-4.44	$< 10^{-4}$
3. bili	0.107	7.02	$< 10^{-4}$
4. edema	0.811	2.63	0.0084
5. hepatom	0.594	2.81	0.0049
6. prothrombin	0.26	3.34	0.0008

Como se puede ver, la variable más significativa resultó ser *bili* que justamente fue aquella que mejor clasifica los pacientes de acuerdo a su pronóstico. Las 5 variables *ascites*, *alk*, *platelet*, *sex* y *spiders*, correspondientes a los coeficientes β_3, β_4, β_8, β_{10} y β_{11} del primer modelo, fueron omitidas en el segundo dado que el cociente de verosimilitud con el primer y segundo modelo

resultó ser:

$$-2(-553.949 + 550.3224) = 7.2532$$

El cual tiene una distribución $\chi^2_{(5)}$ y al comparar con $\chi^2_{(5),0.95} = 11.0705$, se puede ver que el ajuste de ambos modelos no es significativamente distinto, por lo cuál se opta por el modelo más simple del **cuadro 3.20 b)** (principio de parsimonia). Y aunque este modelo se podría considerar como adecuado, podría mejorarse el ajuste con alguna transformación. Se decide aplicar *log()* a algunas variables debido a la suposición que que un cambio en los valores de x a $x + d$ podría generar un gran impacto en $\lambda(t)$ si $x \to 0$. De esta manera, se hace una transformación de las covariables y se agrega al último modelo las covariables transformadas *log(age)*, *log(albumin)*, *log(bili)* y *log(prothrombin)*, luego se observa el cociente de verosimilitudes para evaluar la mejora. En el **cuadro 3.21** se observan los resultados luego de hacer el proceso de selección de variables.

Cuadro 3.21: Estimación de los coeficientes del modelo de riesgos proporcionales ajustado con las variables del **cuadro 3.16** más las transformaciones de *log()* de *age*, *albumin*, *bili* y *prothrombin* correspondientes en cada paso del algoritmo de selección de la función *stepAIC()* con los datos de la base *pbc* con los 312 que recibieron DPCA.

a) *log verosim*: -537.9788	*Coef.* $\hat{\beta}_i$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
1. age	-0.027	-0.39	0.6979
2. albumin	1.006	0.59	0.5531
3. bili	-0.045	-1.29	0.1957
4. edema	0.828	2.72	0.0064
5. hepatom	0.195	0.89	0.3746
6. prothrombin	-0.608	-0.53	0.5952
7. log(*age*)	3.224	0.86	0.3919
8. log(*albumin*)	-5.863	-1.09	0.2743
9. log(*bili*)	1.077	5.10	$< 10^{-4}$
10. log(*prothrombin*)	10.19	0.75	0.4503
b) *log verosim*: -541.064	*Coef.* $\hat{\beta}_i$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
1. age	0.033	3.933	$< 10^{-4}$
2. albumin	-0.945	-3.984	$< 10^{-4}$
3. edema	0.804	2.689	0.0071
4. prothrombin	0.246	2.920	0.0035
5. log(bili)	0.886	8.997	$< 10^{-4}$
c) *log verosim*: -540.7804	*Coef.* $\hat{\beta}_i$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
1. age	0.0336	3.881	0.0001
2. edema	0.7881	2.635	0.0084
3. log(albumin)	-3.0488	-4.207	$< 10^{-4}$
4. log(bili)	0.8814	8.925	$< 10^{-4}$
5. log(prothrombin)	3.0137	2.942	0.0032

Comparando los modelos anidados del **cuadro 3.21 a)** y el **cuadro 3.20 b)**, se calculó el cociente de log-verosimilitudes para evaluar qué tan significativa es la diferencia en el ajuste con las

covariables transformadas. Con lo cual se obtiene

$$-2(-553.9490 + 537.9788) = 31.9404$$

Y dado que $\chi^2_{(5),0.95} = 9.4877$, se puede concluir que las transformaciones aportan una gran mejora al modelo. Por otro lado, en los demás modelos a) y c) del **cuadro 3.21** no se tiene una diferencia significativa de los valores de la verosimilitud, de modo que se opta por el modelo más simple del **cuadro 3.17**, pues el cociente de verosimilitud es

$$-2(-540.7804 + 537.9788) = 5.6032$$

el cuál es menor a $\chi^2_{0,95,5} = 11{,}0705$ y por lo tanto, se prefiere el modelo en el **cuadro 3.20 c)** más simple. En la siguiente sección se procede a hacer el análisis de residuales para evaluar las transformaciones del modelo y la influencia que tiene cada una en la predicción del modelo.

Evaluación de las transformaciones

La primera variable a evaluar es *bili* por ser la más significativa en el modelo. Para ello se hace el análisis con un modelo en el que se ajustan las siguientes variables:

$$\lambda(t) = \lambda_0(t) \cdot exp(\beta_1 \cdot age + \beta_2 \cdot \log(albumin) + \beta_3 \cdot \log(prothrombin) + \beta_4 \cdot edema) \tag{3.2.1}$$

Esto con el objeto de evaluar el efecto que tiene la variable antes de ser incorporada al modelo. En la **fig. 3.33** se exponen la gráfica de los residuales del modelo **3.2.1** vs. *log(bili)* (izquierda) y vs. *bili* (derecha). En cada gráfica se ajusta una curva polinomial que describe la tendencia de los residuos. Se puede ver que para la gráfica de los residales vs. *log(bili)* se tienen errores menos irregulares, pues en la gráfica vs. *bili* los errores se acumulan justamente para los valores más pequeños, lo cuál es coherente con el diagnóstico médico mencionado en la sección anterior. Dado que en la gráfica derecha se tiene una distribución más homogénea de los residuales, tiene sentido pensar que la transformación *log()* favorece el ajuste para la variable *bili*.

Adicionalmente, se desea investigar la posibilidad de que la hepatomegalia (crecimiento anormal del hígado) tenga alguna influencia en el ajuste de *bili* o en la transformación *log()* en el modelo. Por ello en la **fig. 3.34** se observa una comparación hecha en una población que no presenta hepatomegalia y en la **fig. 3.35** se puede ver la comparación con la población con hepatomegalia.

Figura 3.33: Residuales de martingala del modelo **3.2.1** con las variables *age*, log(*albumin*), log(*prothrombin*) y *edema* vs. log(*bili*) (izquierda) y vs. *bili* (derecha)

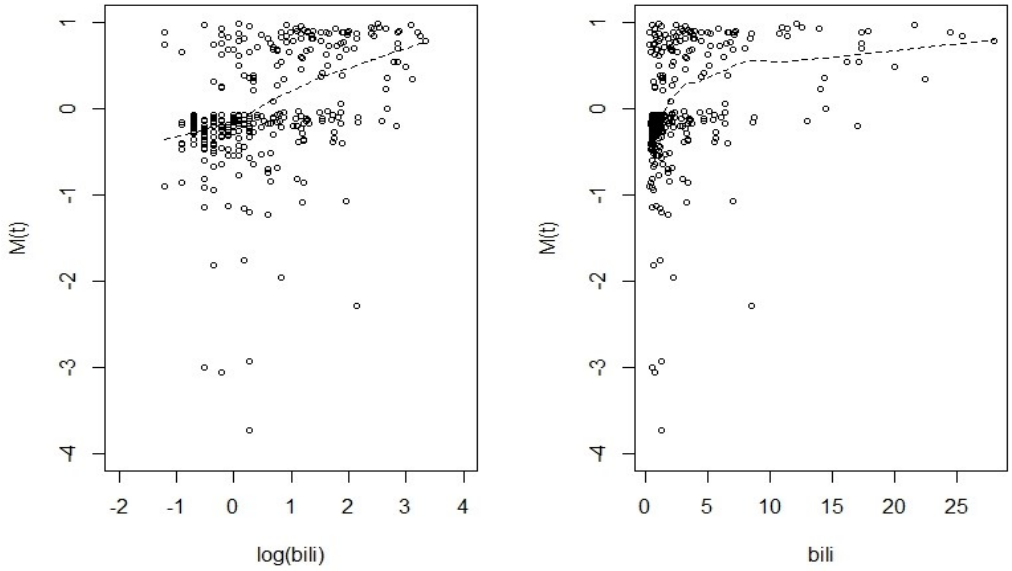
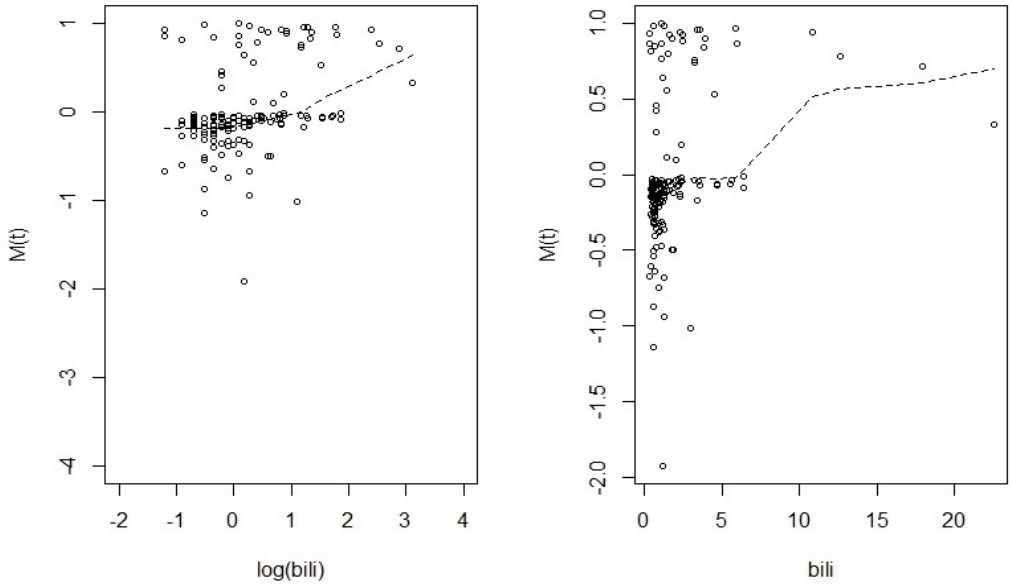
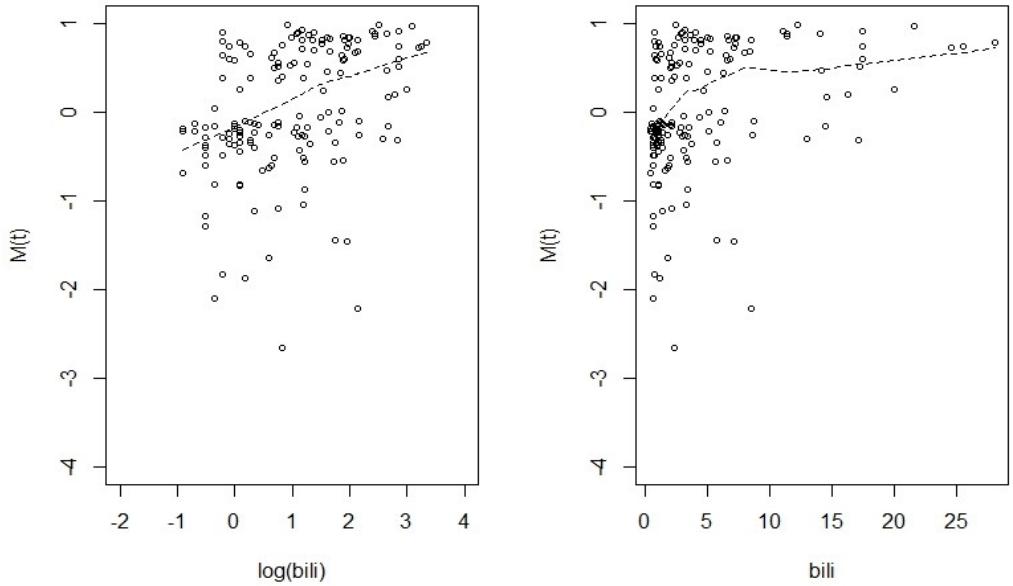


Figura 3.34: Residuales de martingala del modelo **3.2.1** con *age*, log(*albumin*), log(*prothrombin*) y *edema* vs. log(*bili*) y vs. *bili*. Correspondientes al ajuste con 152 observaciones de pacientes sin hepatomegalia.



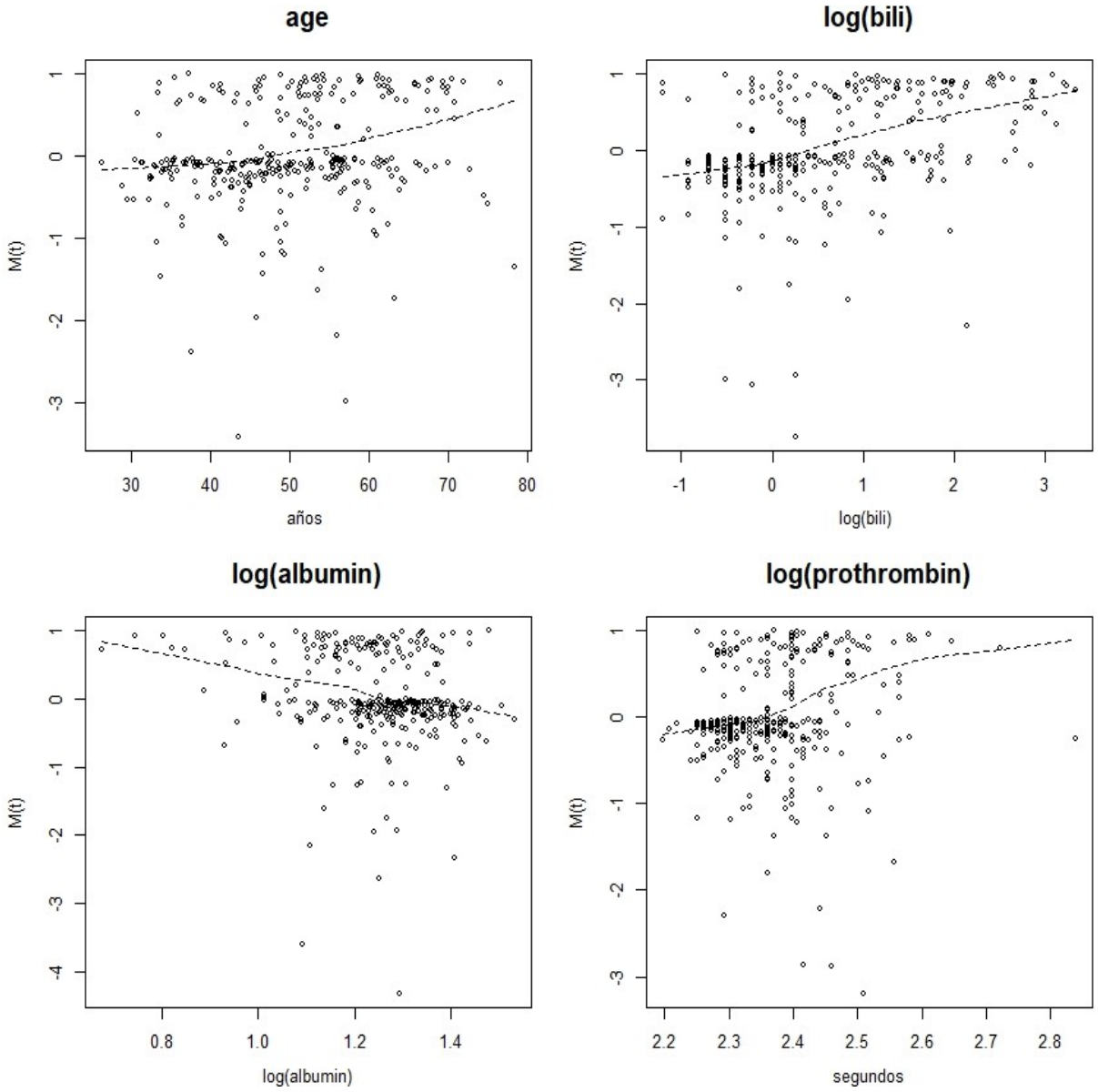
Primeramente se observan las gráficas derechas de la **fig. 3.34** y **fig. 3.35**, donde se grafican los residuales vs. log(*bili*) y lo que se ve es que realmente no hay una diferencia significativa, por lo cual se puede concluir que la hepatomegalia no tiene una influencia en la variable transformada log(*bili*). Si ahora se observan las gráficas izquierdas de la **fig. 3.34** y **fig. 3.35**, se puede ver que la tendencia en los primeros valores de log(*bili*) (los valores < 7) es considerablemente distinta al resto, así se observa que la hepatomegalia influye localmente en los valores pequeños de *bili*.

Figura 3.35: Residuales del modelo **3.2.1** con *age*, log(*albumin*), log(*prothrombin*) y *edema* vs. log(*bili*) y vs. *bili*. Correspondientes al ajuste con 160 observaciones de pacientes con hepatomegalia.



Para evaluar el ajuste de las demás variables, se trabaja con las mismas variables *age*, *edema* y las transformaciones log(*albumin*), *log(bili)* y *log(protime)*. Se grafican los residuales de los modelos que contengan todas las variables exceptuando la que se quiere analizar (que es aquella que aparecerá en el eje de las abscisas). Los resultados se muestran en las gráficas de la **fig. 3.36**.

Figura 3.36: Residuales del modelo con cuatro de las covariables *age*, log(*albumin*), log(*protime*), *edema* y log(*bili*) graficados vs. la variable omitida (aquella que se omite en el modelo y aparece en el eje horizontal)



Al contemplar las gráficas, se pueden evaluar las transformaciones de las variables y al respecto, se puede decir que no hay un problema significativo con ellas ya que los errores parecen ser homogéneos a lo largo de los distintos valores de las covariables, que era justamente lo que no sucedía con el primer análisis de la covariable *bili*. Finalmente, se considera que nuestro modelo está dado por la expresión:

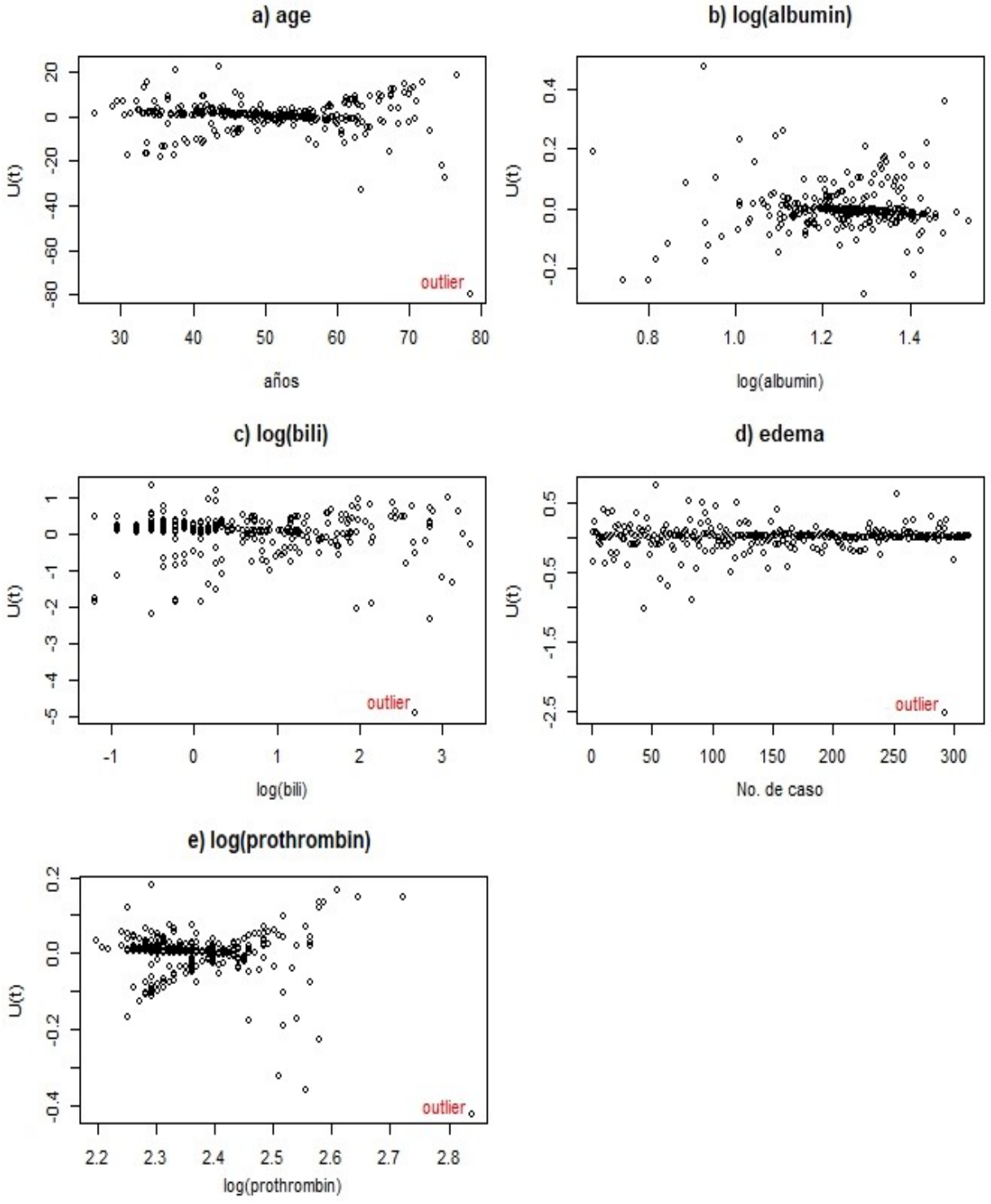
$$\lambda(t) = \lambda_0(t) \cdot exp(\beta_1 \cdot age + \beta_2 \cdot edema + \beta_3 \cdot \log(albumin) + \\ \beta_4 \cdot \log(bili) + \beta_5 \cdot \log(prothrombin)) \tag{3.2.2}$$

Búsqueda de puntos de influencia

En esta sección se analizarán los residuales de score que corresponden al modelo **3.2.2** con las 5 covariables. Se grafica respecto a cada covariable para buscar valores extremos que hagan sospechar sobre un posible punto de influencia o algún error en la medición de las covariables. Para ello se tienen los resultados resumidos en la **fig. 3.37**, donde se grafican los residuales de score de cada covariable con respecto a su correspondiente variable. Se realiza entonces una búsqueda de puntos que podrían influenciar el modelo y lo que atrae la atención en la **fig. 3.37** son las gráficas correspondientes a log(*bili*) y *edema*, pues en ellas se ven puntos muy alejados del resto (señalados como outliers). Estos puntos corresponden, de acuerdo con *Fleming & Harrington*, a errores de captura de los datos. Como el caso del paciente #81: una mujer de 64 años sin edema, buena cantidad de albúmina y un alto tiempo de coagulación sanguínea. A pesar de tener una gran cantidad de bilirubina, su tiempo de supervivencia fue bastante largo. Por ello podría pensarse que se trata de un error en la captura de sus datos.

Otros casos que más llamaron la atención fueron del paciente # 253, que corresponde al del paciente más viejo del estudio (véase **fig. 3.37**, gráfico de *age*) y del paciente # 107 con el tiempo de coagulación más largo (véase gráfico de *log(prothrombin*). Con base a los diagnósticos de ambos pacientes se pudo observar lo siguiente: dado que el paciente #253 tenía un alto nivel de bilirubina y un bajo nivel de albúmina se esperaba que viviera poco tiempo, mientras que el paciente # 107 también vivió demasiado tiempo a pensar de tener mala coagulación.
Los perfiles inusuales de estos pacientes provocaron incertidumbre en la calidad de los datos y luego de una revisión se corroboró que los datos estuvieron mal capturados (el paciente más viejo en realidad tenía 54 años en lugar de 78 y el otro paciente tenía un tiempo de coagulación de 10.7 segs. en lugar de 17.1). Esta fue la manera en la que se detectaron irregularidades en los datos y que se pudieron contrastar los datos errados de aquellos que sólo eran atípicos, los detalles se encuetran en *Fleming & Harrington* (1991, pág 184).

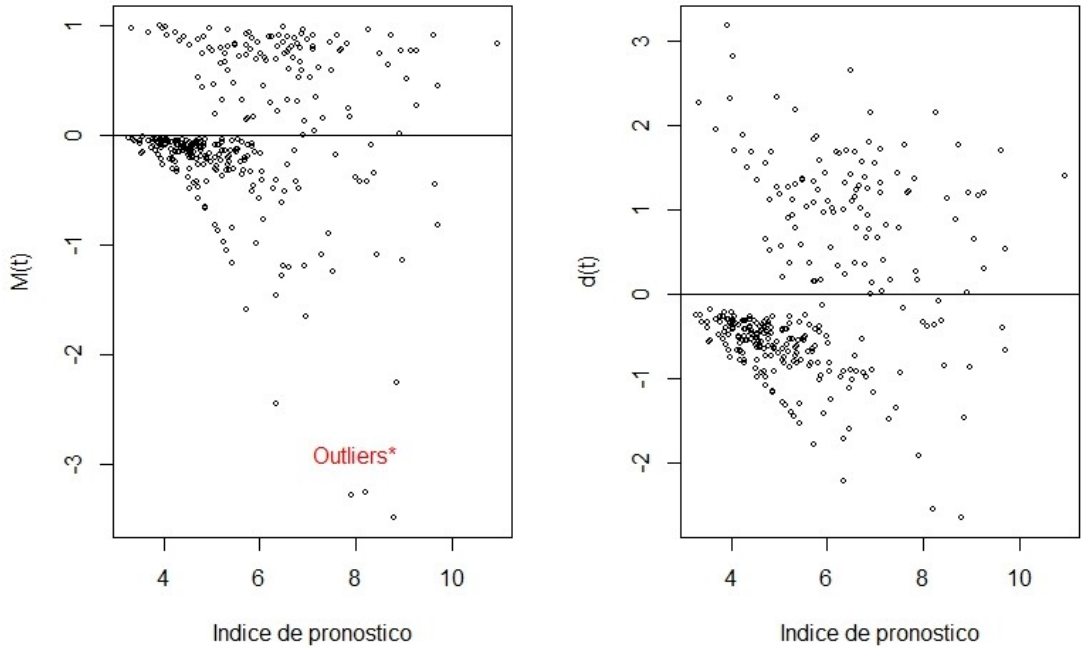
Figura 3.37: Residuales de score del modelo **3.2.2** con las covariables *age*, log(*albumin*), log(*prothrombin*), *edema* y log(*bili*) graficados vs. su correspondiente variable a analizar. Los residuales de *edema* se grafican vs. el número de caso.



Evaluación de la exactitud del modelo

Para observar los posibles outliers considerando todas las variables en conjunto, se muestran en la siguiente figura los residuales de martingala y devianza (izquierda y derecha respectivamente) del modelo **3.2.2**, graficados vs. el índice de pronóstico calculado con el mismo modelo. Al graficar los residuales vs. el índice de pronóstico se buscan nuevamente valores atípicos, los cuales parecen existir (se muestran a la izquierda), pero al observar los residuales de devianza (derecha), se concluye que finalmente estos valores no son atípicos. He aquí algunos de los beneficios de hacer simétricos los residuales.

Figura 3.38: Residuos de martingala con el modelo **3.2.2** vs. *índice de pronóstico* (izquierda) y residuos de devianza (derecha) vs. *índice de pronóstico*.



Validación del supuesto de proporcionalidad

Para corroborar que las variables cumplen con el supuesto de ser proporcionales en el modelo, se procede a analizar las curvas estratificadas de cada variable (las variables continuas se dividieron en cuartiles) bajo la transformación $log(-log(S(t))$ de la función de riesgo de base del modelo obtenida de la estimación de Kaplan-Meier, tales curvas se espera que sean paralelas. La **fig. 3.39** muestra los resultado de las curvas. Lo que se puede observar en dichas gráficas es que se tienen curvas aproximadamente paralelas en todas las covariables, a excepción de *edema* y log(*prothrombin*). Además, se prueba la hipótesis de proporcionalidad para la j-ésima variable:

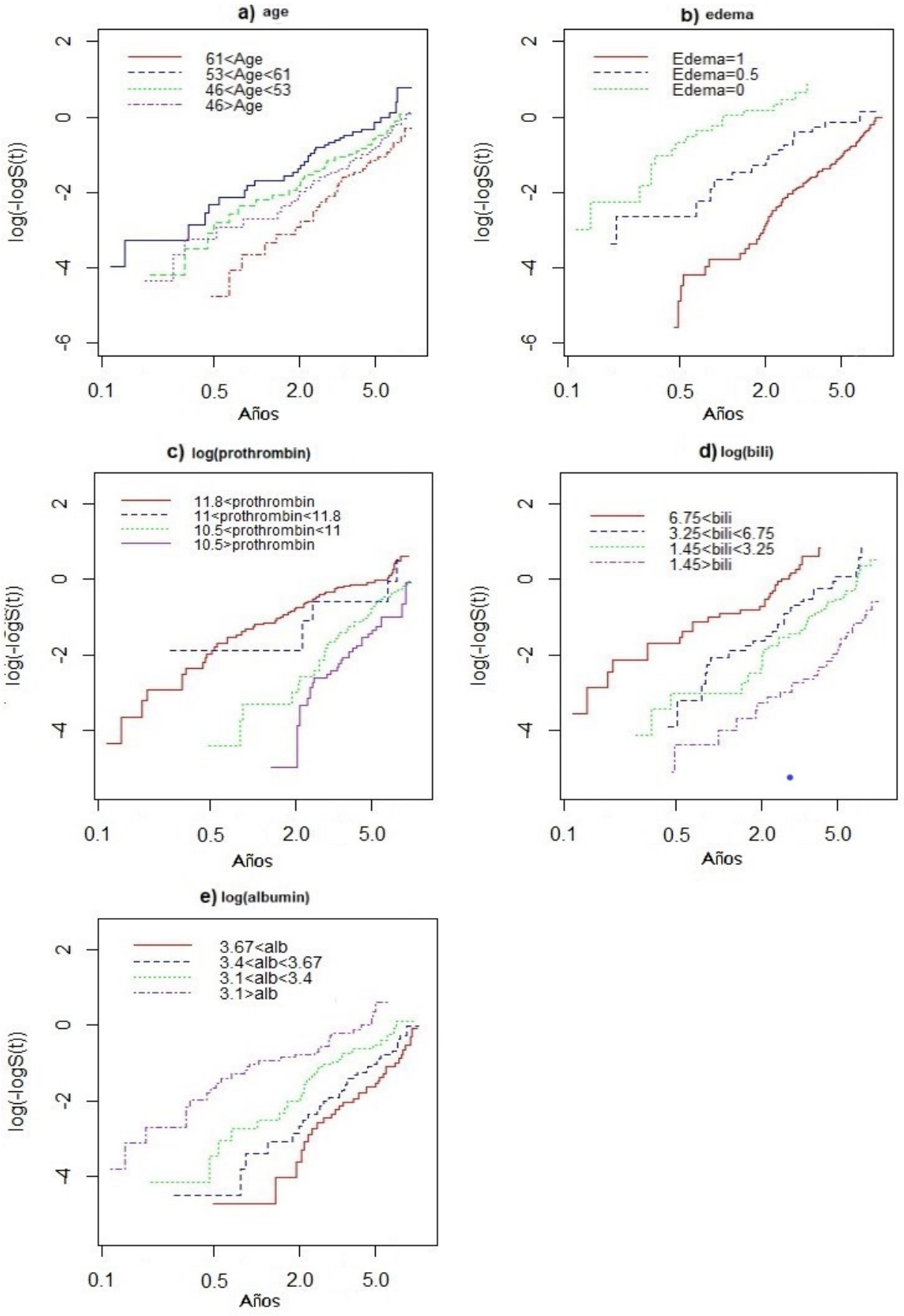
$$H_0 : \lambda(t) = \lambda_0(t)exp(\beta_j X_j) \text{ vs. } H_1 : \lambda(t) \neq \lambda_0(t)exp(\beta_j X_j)$$

Junto con el apoyo gráfico, se muestran los resultados de la prueba formal de la hipótesis anterior dados por la función *cox.zph()* de R en el siguiente cuadro.

Cuadro 3.22: Prueba de proporcionalidad para cada variable del modelo **3.2.2**. La columna *chisq* representa el valor de la estadística de prueba que se distribuye $\chi^2_{(5)}$.

Variables	*chisq*	*P-valor*
age	0.14	0.70
log(*albumin*)	0.05	0.82
log(*bili*)	2.33	0.13
edema	2.65	0.10
log(*prothrombin*)	3.25	0.07
GLOBAL	8.93	0.11

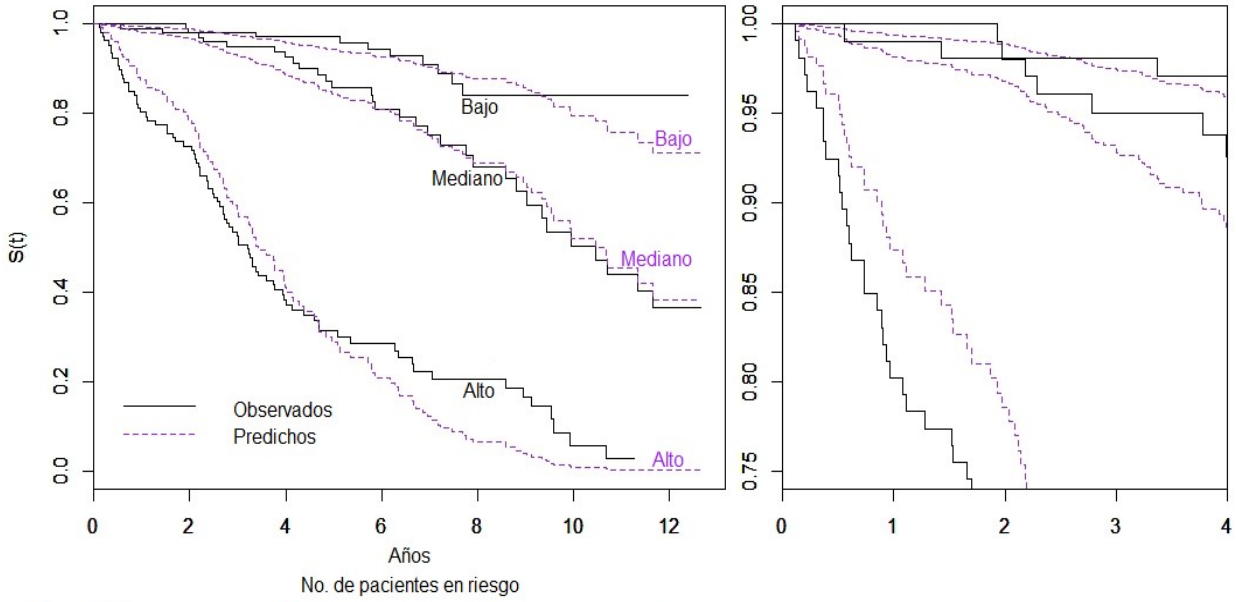
Figura 3.39: Estimador de Kaplan-Meier estratificadas según sus cuartiles (para variables continuas) o según sus niveles (para *edema*) bajo la transformación $log(-log(S(t))$ del modelo **3.2.2**.



Grupos de riesgo

Para realizar el análisis, se empleará el Índice de pronóstico como medida del riesgo en el diagnóstico de los pacientes. De este modo, se ocupa el modelo **3.2.2** del que se tienen 3 grupos definidos por los terciles del Índice de pronóstico, de tal manera que se puede estratificar la población con esta nueva variable. En la **fig. 3.40** se tienen representados el estimador de Kaplan-Meier estratificado para cada grupo de riesgo definido con el Índice de pronóstico, estas curvas representan las observaciones de los pacientes. Adicionalmente, en color azul se tienen las curvas de la estimación de la supervivencia $S(t)$ evaluadas en los valores definidos en el **cuadro 3.23**.

Figura 3.40: Estimaciones de la función de supervivencia evaluadas en la media de cada variable de cada grupo de riesgo (azul) junto con las observaciones del grupo correspondiente representadas por el estimador de Kaplan-Meier estratificado por los grupos de riesgo. En la parte inferior se muestran la cantidad de individuos en riesgo en los tiempos indicados.



Cuadro 3.23: Valores de las covariables en la que se evalúa la función de supervivencia del modelo para cada grupo de riesgo.

	Bajo	*Medio*	*Alto*
age	46	56.34	53.03
edema	0.018	0.067	0.129
log(*bili*)	-0.187	0.285	1.833
log(*albumin*)	1.337	1.215	1.129
log(*prothrombin*)	2.345	2.349	2.437

También hay que mencionar que luego de realizar la prueba de log rank se obtiene un *p-valor*$\to 0$, por lo que no se rechaza la hipótesis de que las estratificaciones de grupos de riesgo son distintas. De este modo, se tiene una buena clasificación.

Ajuste de modelo de riesgos propocionales no paramétrico

Se ajusta el siguiente modelo de riesgos proporcionales no paramétrico, con las covariables del modelo **3.2.2**:

$$\lambda(t) = exp(\beta_0(t) + \beta_1(t) \cdot age + \beta_2(t) \cdot edema + \beta_3(t) \cdot \log(albumin) + \\ \beta_4(t) \cdot \log(bili) + \beta_5(t) \cdot \log(prothrombin)) \tag{3.2.3}$$

Con este análisis, se desea probar la hipótesis de que las variables tienen un efecto constante a lo largo del tiempo, i.e. se desea entonces contrastar, para cada j-ésima variable, la hipótesis de que sus correspondientes coeficientes de regresión son constantes, o dicho de otro modo:

$$H_0 : B_j(t) = 0 \text{ vs. } H_1 : B_j(t) = \beta_j$$

La forma en la que se realiza con la prueba de Kolmogorov-Smirnov, cuyos resultados son los siguientes:

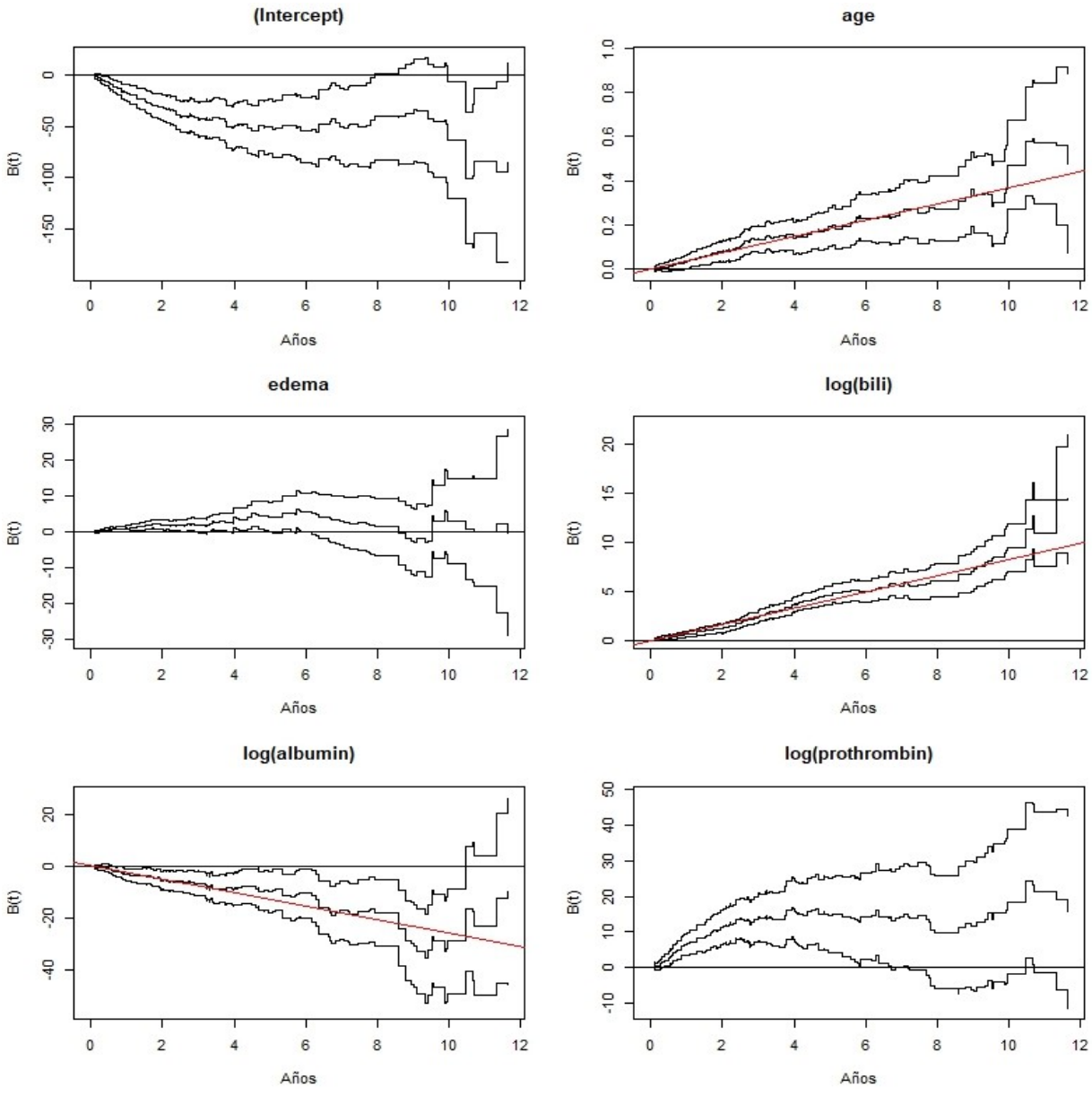
Cuadro 3.24: Prueba de hipótesis de efecto constante de las covariables del modelo **3.2.3** basada en la estadística $B_j(t) - (\frac{t}{\tau})B_j(\tau)$ para $t \in [0, \tau]$.

	$sup\|B_j(t) - (t/\tau)B_j(\tau)\|$	*P-Valor*
intercept	2.068	$< 10^{-4}$
age	0.041	0.96
edema	5.367	$< 10^{-4}$
log(*bili*)	0.716	0.65
log(*albumin*)	3.534	0.786
log(*prothrombin*)	12.83	$< 10^{-4}$

En la tabla se puede observar conforme a los p-valores, que las variables que significativamente no son de efecto constante son *edema* y log(*prothrombin*). Esto es interesante porque resultaron

ser las variables que no cumplieron con el supuesto de proporcionalidad en el modelo anterior paramétrico. En la **fig. 3.41** se observa una gráfica del coeficiente de regresión acumulado estimado $\hat{B}(t)$ vs. t de las covariables.

Figura 3.41: Estimaciones de los coeficientes acumulados para las variables del modelo **3.2.3** no paramétrico. Para aquellos covariables que significativamente tienen un efecto constante a lo largo del tiempo, se muestra una recta cuya pendiente representa el coeficiente estimado.



Luego de haber determinado qué variables tienen significativamente efecto constante de aquellas que no, se puede ajustar un modelo semi-paramétrico más simple, en el cuál no todas las variables se presumen ser de efecto variante, sino que algunas de ellas se representan con efecto constante. Por ello, el modelo semi-paramétrico que finalmente se ajusta quedará determinado por *edema* y log(*prothrombin*) como variables de efecto variante y log(*albumin*), log(*bili*) y *age* como

variables de efecto constante. De tal modo, se tendrá el siguiente modelo:

$$\lambda(t) = exp(\beta_0(t) + \beta_1(t) \cdot edema + \beta_2(t) \cdot \log(prothrombin) + \gamma_1 \cdot age + \gamma_2 \cdot \log(bili) \\ +\gamma_3 \cdot \log(albumin) + \gamma_4 \cdot \log(bili) + \gamma_5 \cdot \log(prothrombin)) \tag{3.2.4}$$

Una vez que se determinan las variables con efecto constante en el tiempo, los resultados de las estimaciones de los coeficientes pueden observarse en el siguiente cuadro:

Cuadro 3.25: Coeficientes del modelo semi-paramétrico **3.2.4** para las covariables *age*, log(*albumin*) y log(*bili*).

Variable	*coef.*$\hat{\beta}_i$	$\hat{\beta}_i/\sigma_i$	*P-valor*
age	0.037	4.07	$< 10^{-4}$
log(*bili*)	0.834	8.35	$< 10^{-4}$
log(*albumin*)	-2.431	-3.59	$< 10^{-4}$

Prueba de proporcionalidad

Se muestra otra forma de investigar sobre las posibles desviaciones del supuesto de proporcionalidad. Para ello, se tienen 50 realizaciones del proceso de scores (*score process*), denotado como $U_j(\hat{\beta}, t)$ para $j = 1, ..., p$, bajo los supuestos del modelo proporcional para comparar con el proceso evaluado en los datos de la muestra. Las curvas se muestran en la **fig. 3.42**. Adicionalmente, se muestra el *P-valor* para la prueba de hipótesis

$$H_0 : U_j(\hat{\beta}, t) = 0 \text{ vs. } H_1 : U_j(\hat{\beta}, t) \neq 0$$

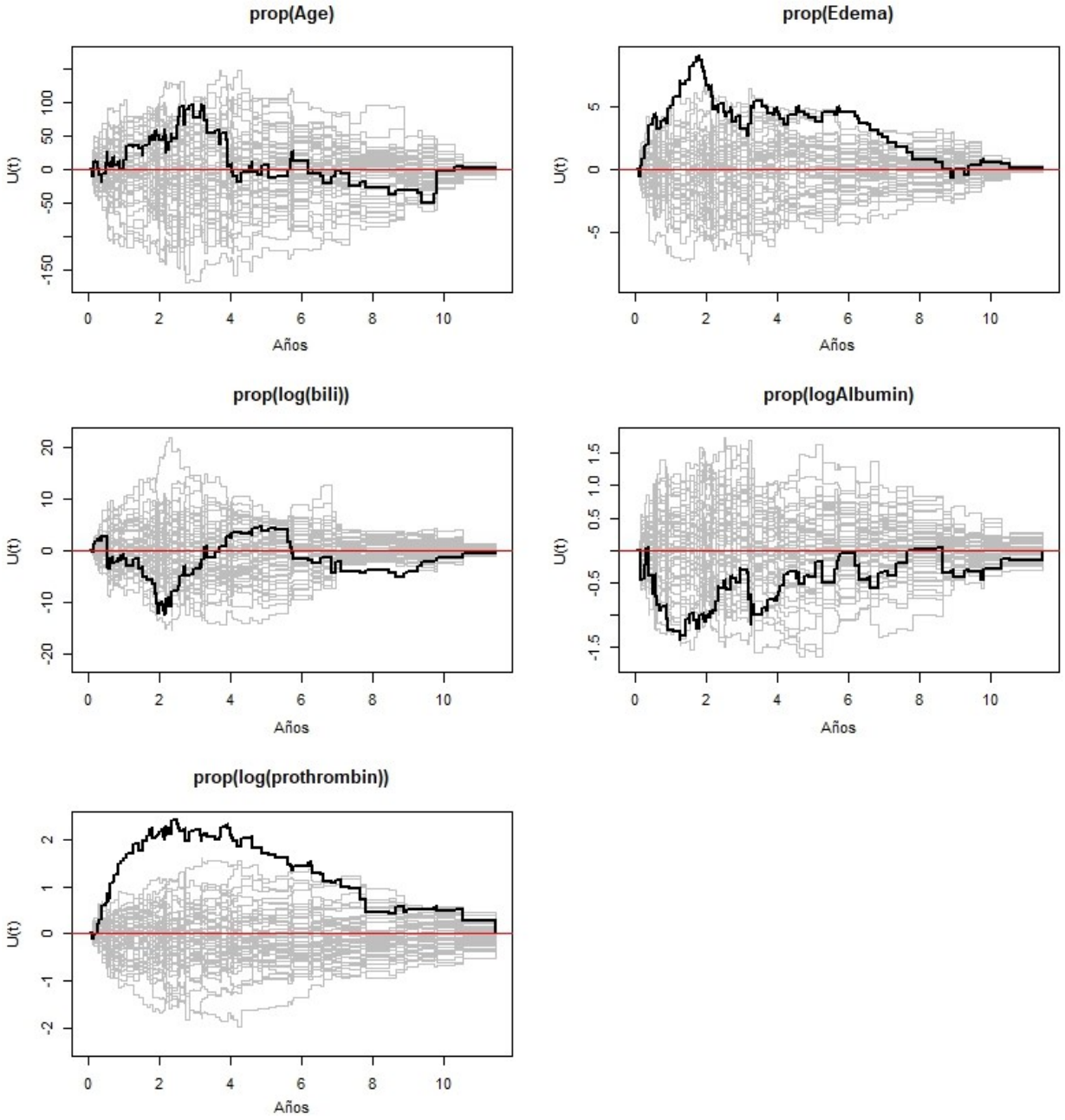
donde la estadística de prueba es:

$$\sup_{t \in [\delta, \tau - \delta]} |U_j(\hat{\beta}, t)|$$

Cuadro 3.26: Prueba de hipótesis de proporcionalidad con el proceso de scores (*score process*) para cada covariable del modelo **3.2.3**

	$sup\|U(t)\|$	*P-valor*
age	112.00	0.315
edema	10.95	0.003
log(*bili*)	12.60	0.154
log(*albumin*)	1.40	0.314
log(*prothrombin*)	2.41	10^{-4}

Figura 3.42: 50 realizaciones del *score process* junto con el proceso obtenido de las observaciones del modelo ajustado **3.2.3** (línea gruesa), desplegado para cada una de las covariables del modelo.



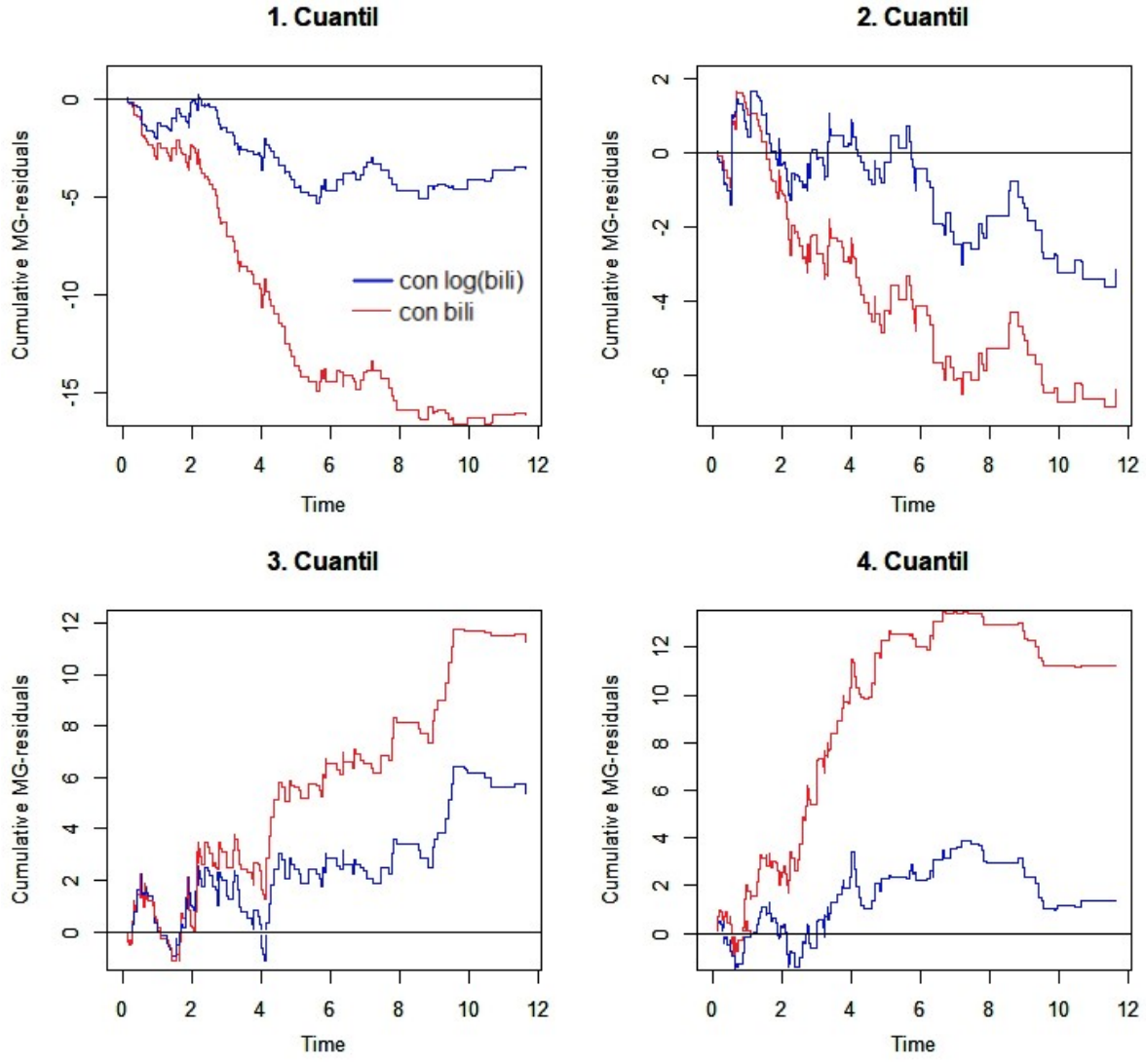
Se observa que existen importantes desviaciones de la curva obtenida del modelo ajustado con respecto a las simulaciones bajo la hipótesis nula para las variables *edema* y log(*prothrombin*), además de que en el **cuadro 3.26** también se muestra que se rechaza la hipótesis de proporcionalidad con un nivel de significancia del 5 % para las mismas variables. Por lo tanto, estas dos variables se consideran como no proporcionales al riesgo en el modelo.

Prueba de bondad de ajuste

Se estiman los residuales de martingalas acumulativos de *bili* para observar su comportamiento y determinar las posibles discrepancias entre $N(t)$ y $\Lambda(t)$ que puedan existir en el modelo ocasionadas por la transformación de $log()$ en esta variable. Para ello, se divide en estratos determinados por los cuartiles y se grafican los residuales del modelo vs el tiempo, así se puede ver el efecto en cada cuartil de la variable en el ajuste. A continuación, se muestra una comparación con y sin aplicar $log()$ a la variable *bili*, del modelo **3.2.3** con el siguiente modelo:

$$\lambda(t) = exp(\beta_0(t) + \beta_1(t) \cdot age + \beta_2(t) \cdot edema + \beta_3(t) \cdot \log(albumin) + \\ \beta_4(t) \cdot bili + \beta_5(t) \cdot \log(prothrombin)) \tag{3.2.5}$$

Figura 3.43: Comparación de las estimaciones de los residuales de martingalas acumulativos estratificados con los cuartiles de la variable *bili* bajo la transformación $log()$ (azul) y sin ella (rojo) en el modelo **3.2.3** y **3.2.5** respectivamente.



Se muestra que el efecto de *log(bili)* parece describir bien el modelo de riesgos proporcionales en comparación con el modelo **3.2.5** con la covariable *bili* sin aplicársele $log()$. El siguiente cuadro muestra una comparación de la prueba de hipótesis de los residuales $H_0 : M_{K_p}(t) = 0$ para cada estrato p de bili correspondiente al modelo **3.2.3** con y sin la transformación de $log()$.

Luego de observar los resultados de las pruebas, se puede concluir que el modelo con *log(bili)* tiene un ajuste satisfactorio, por lo que el modelo de riesgos proporcionales semiparamétrico **3.2.4**, que contiene esa transformación, se considera el más adecuado. Sin embargo, se desea investigar un poco más en otros modelos (aditivo y aditivo-multiplicativo) con el objetivo de ver las mejoras que estos pueden aportar a la descripción de los datos.

Cuadro 3.29: Comparación del contraste de hipótesis $H_0 : M_{K_p}(t) = 0$ para el modelo **3.2.3** y **3.2.5** con y sin la transformación $log()$ en la covariable *bili*.

	Cuartil	$sup\|\hat{M}_{K_j}(t)/SD\|$	*P-Valor*
Con *log()*	1ro	5.30	0.47
	2do	3.60	0.68
	3ro	6.41	0.44
	4to	3.92	0.57
Sin *log()*	1ro	16.61	$< 10^{-3}$
	2do	6.81	0.21
	3ro	11.78	0.03
	4to	13.66	$< 10^{-3}$

Modelo aditivo

En este caso, se hará énfasis una vez más en describir el efecto temporal que tienen las variables de la base *pbc*, si es que lo tienen en el modelo aditivo, y evaluar la bondad del ajuste como anteriormente se ha hecho. Con este fin, se ajusta el modelo de riesgos aditivos de en el que se muestran pruebas para contrastar la hipótesis de significancia y efecto constante de las variables. Primeramente se ajusta el modelo con la variable *bili* sin transformar con el fin de evaluar ese efecto. Es decir, se ajusta el siguiente modelo:

$$\begin{aligned}\lambda(t) = \beta_0(t) + \beta_1(t) \cdot age + \beta_2(t) \cdot edema + \beta_3(t) \cdot bili + \\ \beta_4(t) \cdot \log(albumin) + \beta_5(t) \cdot \log(prothrombin)\end{aligned} \tag{3.2.6}$$

Además, al realizar la prueba de significancia para contrastar las hipótesis

$$H_0 : \beta_j(t) = 0 \text{ vs. } H_1 : \beta_j(t) \neq 0$$

los resultados se resumen en el siguiente cuadro:

Cuadro 3.30: Prueba de hipótesis de significancia de las covariables del modelo **3.2.6**.

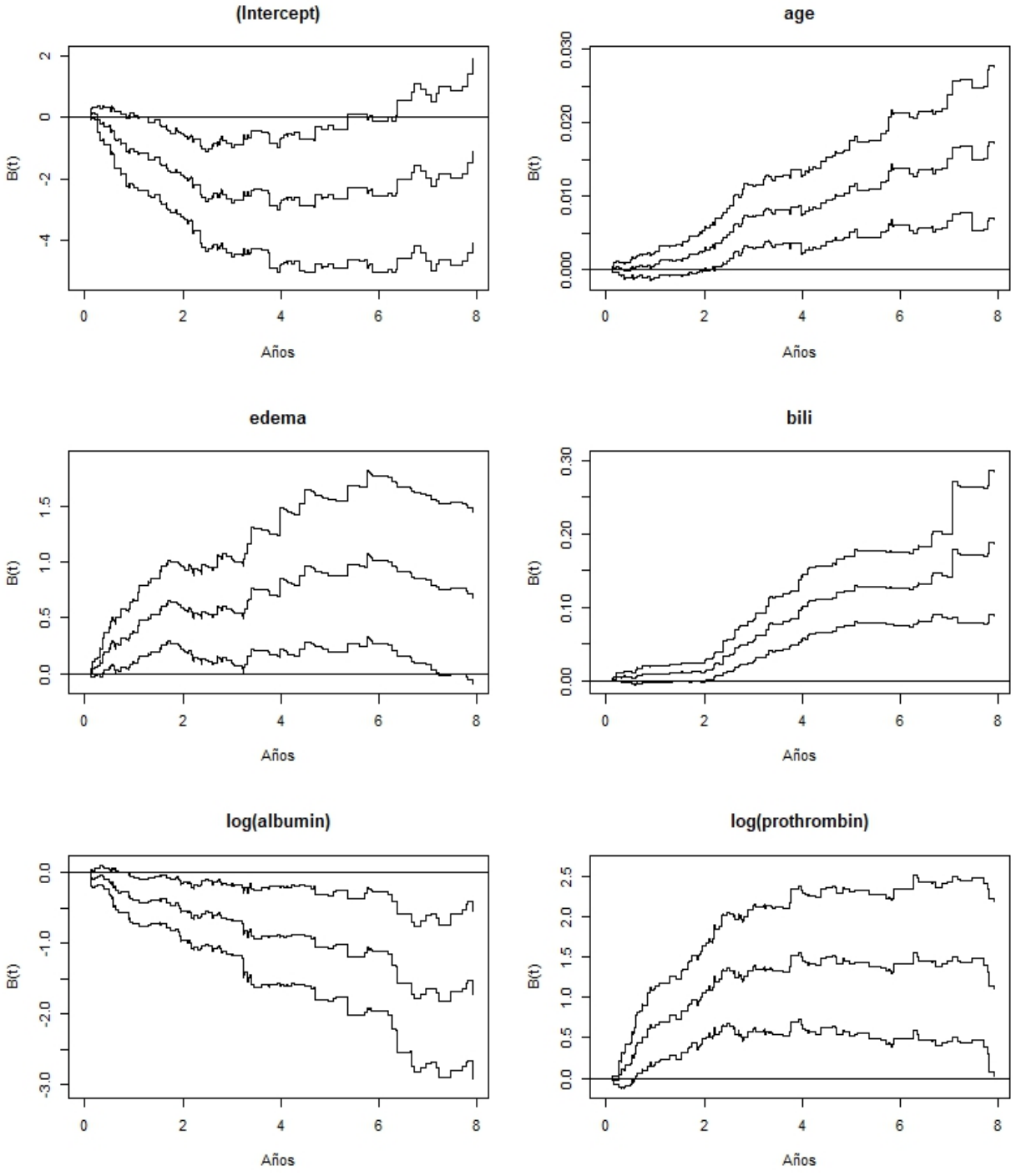
	$sup\lvert\hat{B}_j(t)/\Phi(t)\rvert$	*P-valor*
intercepto	8.43	$< \times 10^{-3}$
age	3.51	0.015
edema	3.17	0.003
bili	5.66	0.001
log(albumin)	3.90	0.011
log(prothrombin)	3.17	0.031

Así pues, esta prueba indica que para todas las variables se rechaza la hipótesis nula y todas son, al igual que en los otros ajustes, estadísticamente significantes. Lo siguiente es observar si tienen un efecto constante en $\lambda(t)$, i.e.

$$H_0 : B_j(t) = \gamma t \text{ vs. } H_1 : B_j(t) \neq \gamma t$$

Como se observa en la columna del *p-valor* en el **cuadro 3.31**, para las variables *age* y *log(albumin)* no se puede rechazar la hipótesis de un efecto constante en el tiempo. La siguiente figura corrobora estas afirmaciones, pues en ella se muestran los valores del coeficiente de regresión acumulado estimado $\hat{B}_j(t)$ para cada variable.

Figura 3.44: Coeficientes de regresión acumulados con las variables *age, edema, bili, log(albumin)* y log(*prothrombin*) del modelo **3.2.6**.



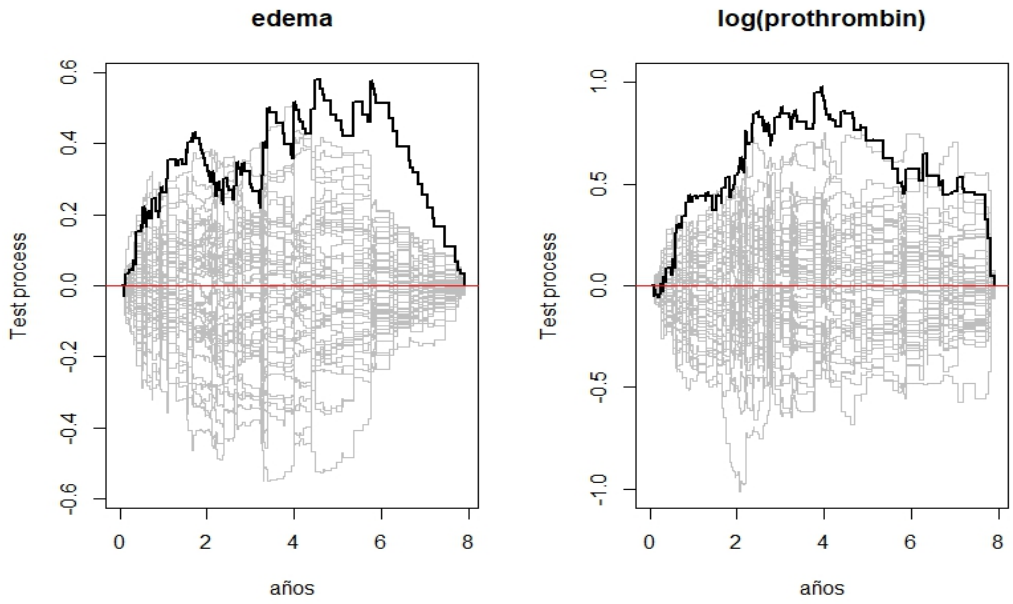
Cuadro 3.31: Prueba de hipótesis Kolmogorov-Smirnov de efecto invariante para cada covariable del modelo **3.2.6**.

	$sup\|B(t) - (t/\tau)B(\tau)\|$	*P-valor*
intercepto	0.150	0.125
age	0.002	0.789
edema	0.429	0.007
bili	0.035	0.318
log(albumin)	0.245	0.791
log(prothrombin)	0.952	0.002

Se puede ver que justamente en la **fig. 3.44**, que en las variables *age* y log(*albumin*), cuyas gráficas se asemejan a las pendientes, se tiene un efecto constante a lo largo del tiempo, que se traduce como un aumento o decremento constante en los coeficientes acumulados de regresión. Vale la pena señalar que si se analiza la curva del efecto que tiene la variable *edema* por intervalos de tiempo, se puede pensar que en los primeros dos años es una variable con efecto constante en el diagnóstico, luego cambia y se mantiene constante casi dos años para finalmente descender.

Además, se tienen las curvas simuladas del *test process* para las covariables *edema* y *log(albumin)*, como otra alternativa para contrastar la hipótesis de efecto constante de las covariables. En la **fig. 3.45** se pueden observar las simulaciones de la prueba.

Figura 3.45: 50 simulaciones del *test process* bajo la hipótesis de efecto constante junto con el proceso del modelo **3.2.6** (línea gruesa) para las variables *edema* y log(*prothrombin*).



Se observan las desviaciones del origen para las variables *edema* y log(*prothrombin*) en comparación de las simulaciones realizadas (en las cuales se supone que tiene un efecto constante). De este modo, con las gráficas de la **fig. 3.45** y la prueba de Kolmogorov-Smirnov, se tiene suficiente evidencia para considerar que sólo *edema* y *log(prothrombin)* son las covariables con efecto variante a lo largo del tiempo y así se puede construir un modelo semi-paramétrico, del cuál se tiene más certeza acerca de la variación del efecto que tienen sus variables en $\lambda(t)$. El modelo que se ajusta es entonces:

$$\begin{aligned}\lambda(t) = \beta_0(t) + \beta_1(t) \cdot edema + \beta_2(t) \cdot \log(prothrombin) + \\ \gamma_1 \cdot age + \gamma_2 \cdot bili + \gamma_3 \cdot \log(albumin)\end{aligned} \tag{3.2.7}$$

Se tienen algunas covariables con efecto constante en el modelo. Los resultados de la estimación de sus coeficientes se muestran en el siguiente cuadro:

Cuadro 3.32: Estimaciones de los coeficientes de las variables con efecto constante del modelo **3.2.7** y prueba de significancia.

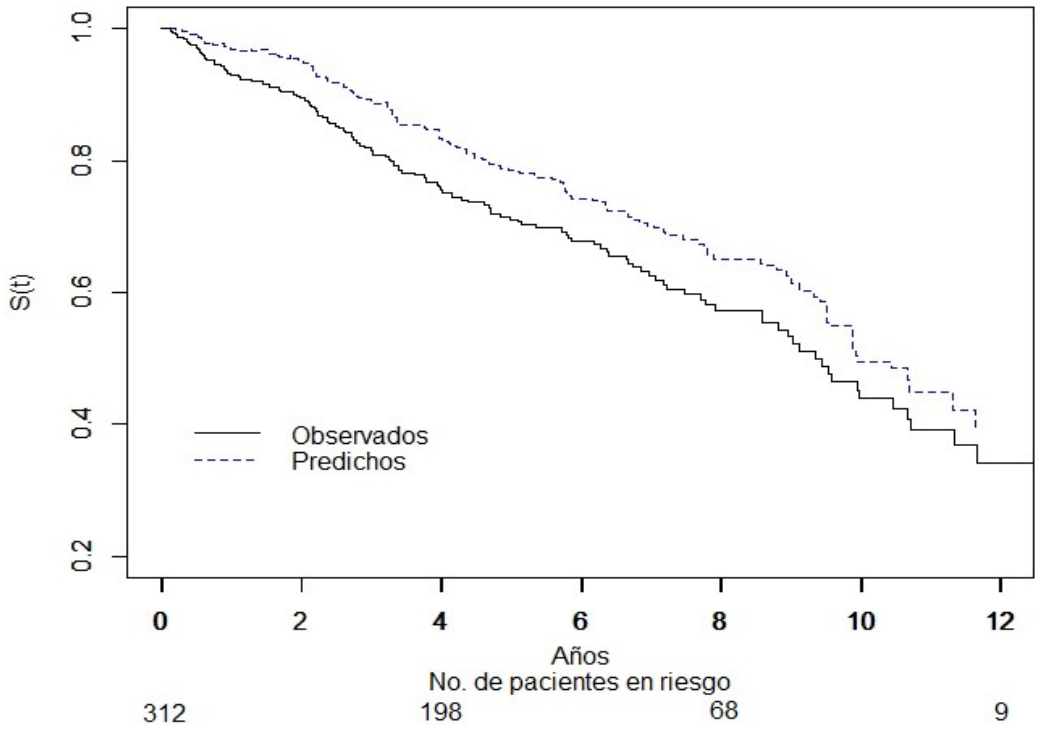
Variable	*Coef.* $\hat{\beta}_i$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
age	0.0020	3.46	0.0005
bili	0.0206	5.35	$< 10^{-8}$
log(*albumin*)	-0.2524	-3.18	0.0009

De acuerdo al cuadro anterior se puede afirmar que el efecto de la variable *age* es pequeño pero significativo, además de que *bili* sin transformar, también lo es como se había visto anteriormente, y asumiendo constante el efecto de *log(albumin)*, se puede asegurar que tiene un efecto negativo en el riesgo del paciente.

Estimación de la función de supervivencia

Una vez determinadas las variables con efecto constante y variante del modelo **3.2.6**, será correcto estimar la función de supervivencia para el modelo aditivo semi-paramétrico. La estimación de la supervivencia se puede observar en la siguiente **fig. 3.46**. La función de supervivencia está evaluada en las medias de cada variable perteneciente al modelo. El ajuste deja qué desear.
Recuérdese que dado que se tienen coeficientes distintos para cada tiempo para las variables *edema* y log(*prothrombin*), no es posible obtener un índice de pronóstico ni elaborar grupos de riesgo.

Figura 3.46: Estimaciones de la función de supervivencia evaluada en las medias de cada variable del modelo semi-paramétrico **3.2.7** como valores predichos. Se compara con la estimación de Kaplan-Meier, como valores observados.



Pruebas de bondad de ajuste

Se procede a analizar los residuales de martingalas acumulativos en distintos estratos definidos por los cuartiles de los datos para la variable *bili*, dado que el modelo se enfoca en analizar el ajuste de esta variable con y sin su transformación. Es decir, se desea comparar el modelo **3.2.6** con el siguiente modelo:

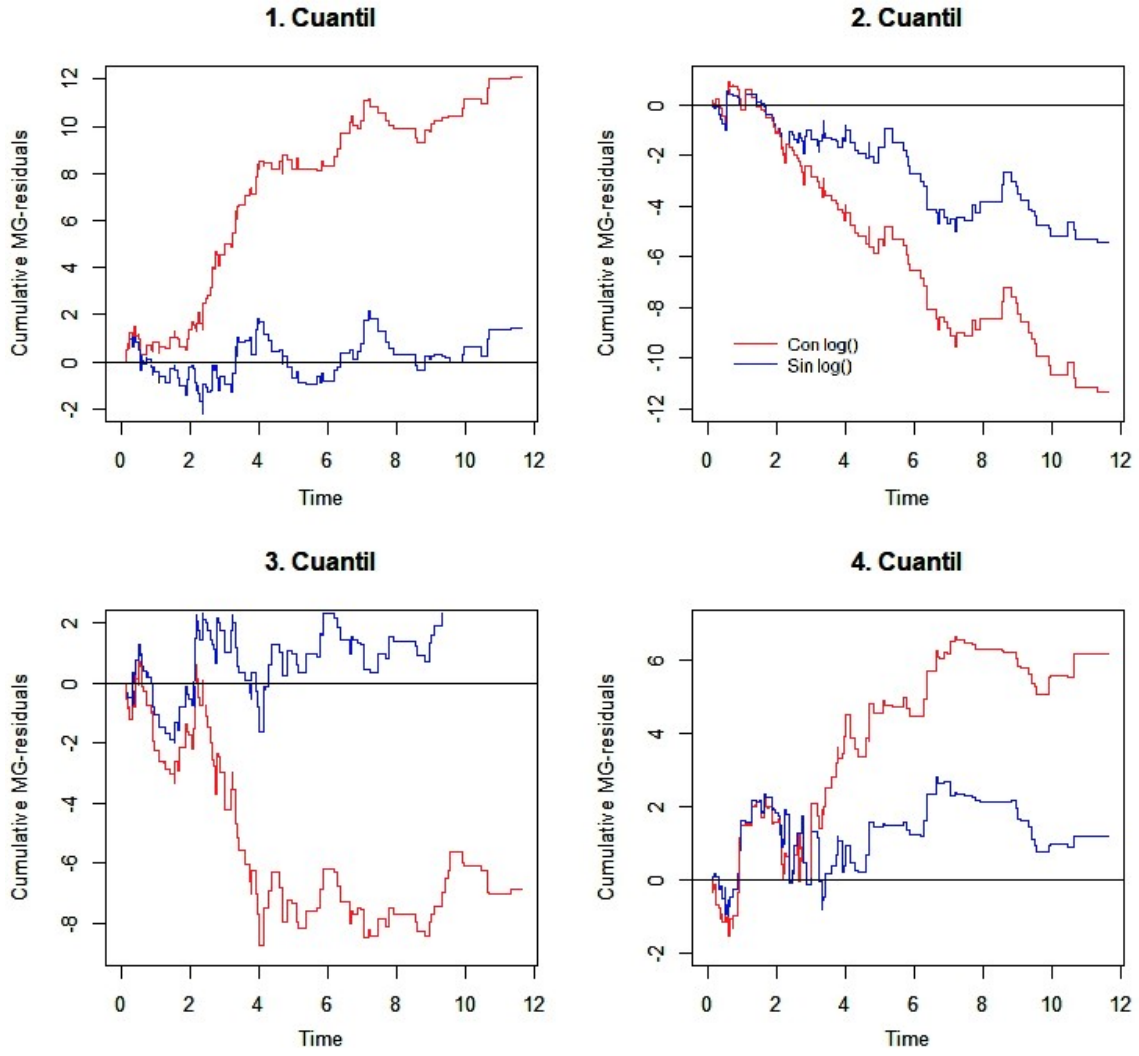
$$\lambda(t) = \beta_0(t) + \beta_1(t) \cdot edema + \beta_2(t) \cdot \log(prothrombin) + \beta_3(t) \cdot age + \\ \beta_4(t) \cdot \log(bili) + \beta_5(t) \cdot \log(albumin) \tag{3.2.8}$$

Como se puede ver en la **fig. 3.47**, se tiene un buen ajuste para el modelo sin la transformación en *bili*, pues a excepción del primer cuartil, los residuales acumulados están al rededor del origen, lo cual indica que las martingalas residuales se aproximan al proceso de conteo. Además se contrasta la siguiente hipótesis, relativa a cada j-ésimo estrato dado por los cuartiles:

$$H_0 : M_{Kj}(t) = 0 \text{ vs. } H_1 : M_{Kj}(t) \neq 0 \tag{3.2.9}$$

Los resultados de la prueba se resumen para los modelos **3.2.6** y **3.2.8** en el siguiente cuadro.

Figura 3.47: Residuales de martingala acumulativos para los cuartiles de la variable *bili* en el modelo **3.2.6** (azul) comparados con los residuales para el modelo **3.2.8** de la covariable log(*bili*) (rojo).



Cuadro 3.29: Comparación del contraste de hipótesis $H_0 : M_{K_p}(t) = 0$ para el modelo con y sin la transformación $log()$ en la covariable *bili*.

	Cuartil	$sup\|\hat{M}_{K_j}(t)/SD\|$	*P-Valor*
Conn *log()*	1ro	12.112	$< 10^{-3}$
	2do	11.365	0.008
	3ro	8.734	0.146
	4to	6.619	0.034
Sin *log()*	1ro	2.187	0.946
	2do	5.431	0.228
	3ro	3.994	0.722
	4to	2.810	0.788

Con ello se puede ver que el ajuste es pobre cuando se usa la transformación del *log()* sobre *bili* en el ajuste del modelo aditivo. Se ajustaba mejor en el caso del modelo de riesgos proporcionales ya que el efecto en el riesgo se daba en incrementos pequeños, y por la manera en que se interpretaban los coeficientes del modelo era mejor realizar así el ajuste. Por lo tanto se considera más adecuado el modelo aditivo semiparamétrico **3.2.6** que no contiene la transformación en *bili*.

Ajuste del modelo aditivo-multiplicativo

Luego de haber realizado los ajustes de los modelos anteriores, se pudo ver que en cada modelo habían variables que no se ajustaban correctamente a los datos con respecto al modelo. Como el caso de de *edema* y *log(prothrombin)* que fueron las variables para las que se rechazó el supuesto de proporcionalidad en el ajuste del modelo. Sin embargo, se tuvieron buenos resultados en el modelo de riesgos aditivos. Por lo tanto, las variables *edema* y *log(prothrombin)* se ajustan al modelo multiplicativo-aditivo de manera aditiva.

Para las demás variables no se rechazó el supuesto de proporcionalidad. De tal manera, *age*, *log(bili)* y *log(albumin)* podrían ajustarse al modelo proporcionalmente. Por lo anterior, la forma en la que se va a modelar la función de riesgo será:

$$\begin{aligned}\lambda(t) = \{\beta_0(t) + \beta_1(t) \cdot edema + \beta_2(t) \cdot \log(prothrombin)\} \\ exp\{\gamma_1 \cdot age + \gamma_2 \cdot log(bili) + \gamma_3 \cdot log(albumin)\}\end{aligned} \tag{3.2.10}$$

A continuación, se muestran las estimaciones hechas para las variables que conforman la parte proporcional.

Cuadro 3.30: Estimaciones y pruebas de proporcionalidad para los elementos proporcionales

	Coef.$\hat{\beta}_i$	$\hat{\beta}_i/\hat{\sigma}_i$	*P-valor*
age	0.0376	0.0034	$< 10^{-4}$
log(*bili*)	0.87	0.0538	0.02
log(*albumin*)	-2.63	0.453	$< 10^{-4}$

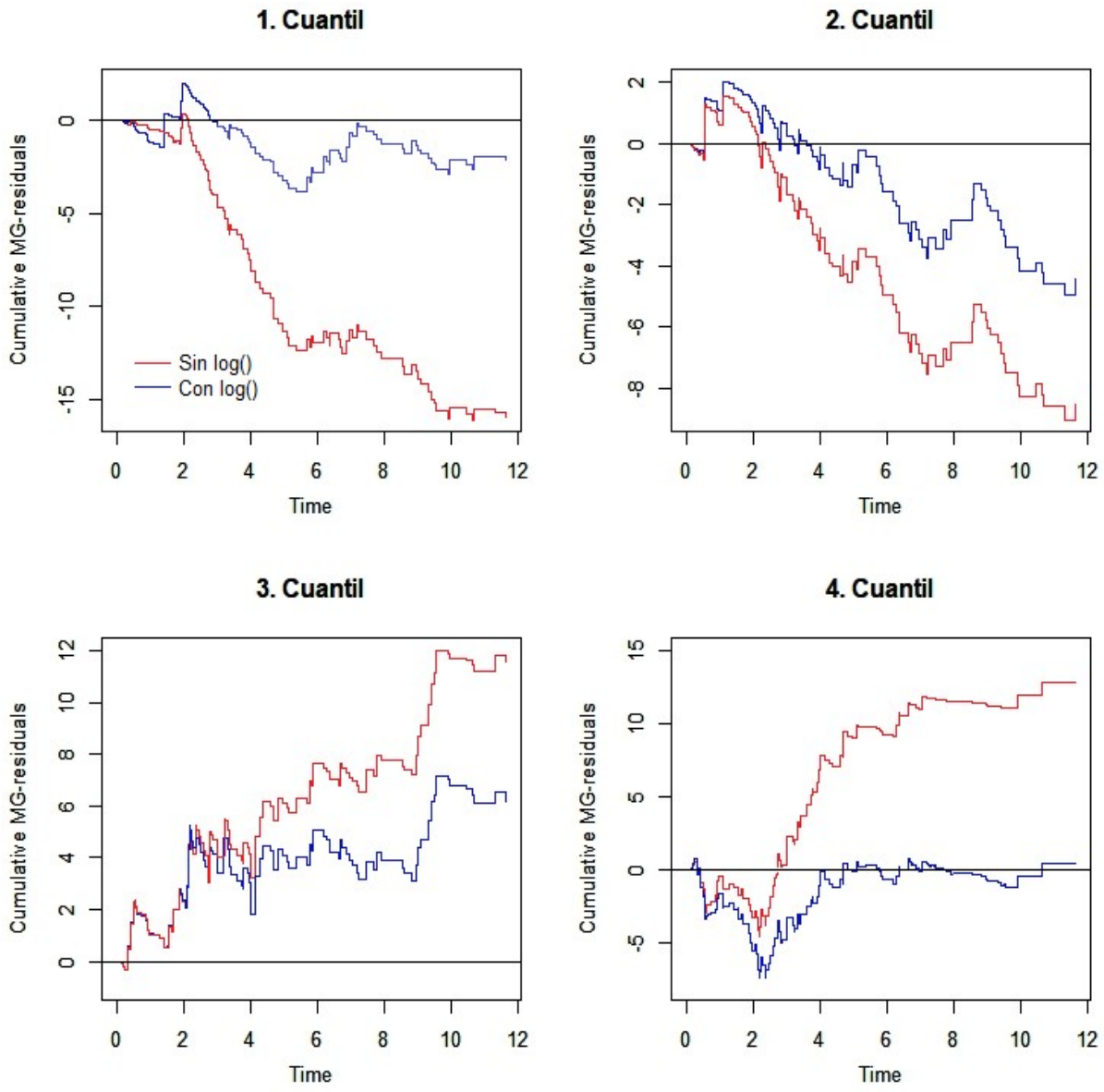
Pruebas de Bondad de Ajuste

Se estiman los residuales de martingalas acumulativos para cada variable y se grafican según los estratos definidos por sus cuartiles o niveles. Se presentan las gráficas correspondientes a la variable *bili* en la **fig. 3.48** ya que, como en la sección anterior, se desea determinar con este criterio si es más preciso usar la transformación de log(*bili*). Por ello, se muestran los residuales de martingala acumulativos para los cuartiles de la variable log(*bili*) del modelo **3.2.10** y *bili* del siguiente modelo:

$$\begin{aligned}\lambda(t) = \{\beta_0(t) + \beta_1(t) \cdot edema + \beta_2(t) \cdot \log(prothrombin)\} \\ exp\{\gamma_1 \cdot age + \gamma_2 \cdot bili + \gamma_3 \cdot log(albumin)\}\end{aligned} \tag{3.2.11}$$

En la siguiente figura se aprecia que los residuales de ambos modelos parecen ser a veces mayores o menores que cero, indicando que a veces el modelo predice fallas de menos o de más. Sin embargo, el modelo con la transformación *log(bili)* parece ser el más adecuado, pues se aleja menos del valor 0.

Figura 3.48: Residuales de martingalas acumulativos para los cuartiles de la covariable *bili* del modelo **3.2.10** y **3.2.11**.



Cuadro 3.29: Coeficientes estimados del modelo aditivo paramétrico **3.1.4**.

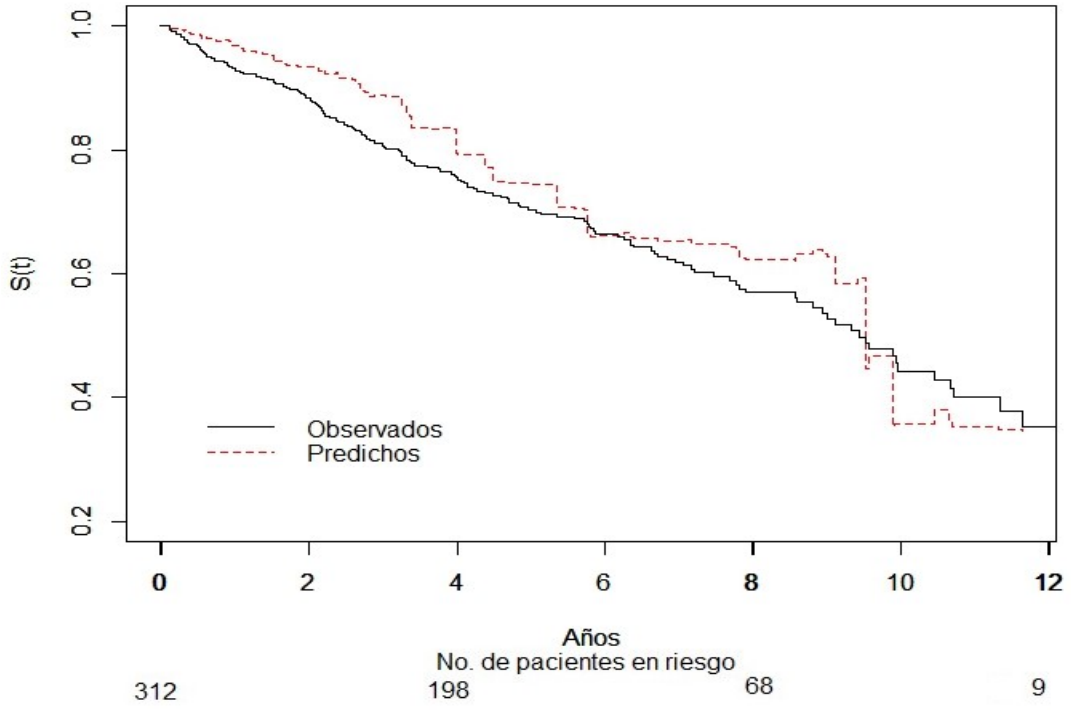
	Cuartil	$sup\|\hat{M}_{K_j}(t)/SD\|$	*P-Valor*
Con *log()*	1ro	3.85	0.29
	2do	4.96	0.26
	3ro	7.19	0.20
	4to	7.39	0.06
Sin *log()*	1ro	16.12	$< 10^{-3}$
	2do	9.05	0.030
	3ro	12.00	0.004
	4to	12.90	$< 10^{-3}$

Al final del análisis, lo que se puede concluir es que el modelo **3.2.10** tiene un mejor ajuste con *log(bili)*, a pesar de que el modelo aditivo **3.2.6** aportaba una mejor predicción sin la transformación de la variable *bili*. Este es sólo un ejemplo en el que se muestra el modelo de riesgos aditivo-multiplicativo como un modelo más flexible y que resulta describir mejor los datos de acuerdo a las hipótesis de proporcionalidad que cumplen las covariables que contiene y la bondad del ajuste que se aprecia al observar sus residuales. Esto sólo pudo verse luego de haber utilizado como herramientas los modelos de riesgo aditivo y proporcional antes de observar los errores en los supuestos que se usaron para ajustar los datos a esos modelo, una vez determinado eso, se pudo saber de qué modo era más correcto modelar las variables de modo que los errores del modelo fueran lo más pequeño posibles.

Estimación de la función de supervivencia

Para las consideraciones médicas, se muestra a continuación la función de supervivencia $S(t)$ estimada por el modelo de riesgos aditivo-multiplicativo evaluada en las medias de cada covariable del modelo, a modo de representar los valores predichos; esto junto con la estimación de Kaplan-Meier para representar los valores observados.

Figura 3.49: Función de supervivencia $S(t)$ estimada la media de cada covariable del modelo **3.2.10** (valores predichos) comparada con la función estimada de Kaplan-Meier (valores observados). Abajo, se muestran los pacientes sobrevivientes en cada tiempo marcado del intervalo de observación.



Capítulo 4

Conclusiones

Para finalizar, no está de más comentar el alcance que se puedo apreciar al desarrollar estos modelos. Se partió de un modelo de riesgos proporcionales que representa una forma sencilla de describir el riesgo a través de un modelo de regresión. La inconsistencia al ajustar este modelo en los estudios de *pbc* y *std* fue que algunas variables no cumplieron el supuesto de proporcionalidad, como se pudo ver el en el ejemplo de *std* para la variable indicadora de clamidia *chla*, mientras que para *pbc* y las variables *prothrombin* (tiempo de producción de protrombina) y *edema*. Por ello no resultaría conveniente aplicar ese modelo. No obstante, en la base *pbc* resultó que aquellas variables no proporcionales, tuvieron un ajuste decente en los otros modelos (aditivo y multiplicativo-aditivo) por lo que fueron excelentes alternativas ante el problema de proporcionalidad de estas variables. Más aún, observar el análisis de residuales de martingala, de score, de devianza y de martingala acumulativos, permitió en ambos estudios, detectar desviaciones del supuesto de proporcionalidad a lo largo del tiempo o en distintos valores de cada variable analizada, proponer transformaciones en las covariables y evaluar la mejora en el ajuste o detectar valores atípicos en las covariables.

Finalmente, otra ventaja de tener flexibilidad al elegir la forma en la que las variables aportan efecto al riesgo en el modelo (ya sea de forma proporcional o aditiva) ha resultado en la forma en la que se clasifican los grupos de riesgo elaborados con el índice de pronóstico. Pues con el ejemplo elaborado en la base *std*, se pudo observar que estos modelos (de riesgos proporcionales, aditivos y aditivos-multiplicativos), brindaron opciones para elegir una clasificación adecuada. En otros casos, esto se puede aplicar de tal manera que un especialista puede utilizar su criterio para la elección del modelo y los grupos de riesgo más adecuados en su estudio.
Como temas de interés fuera del alcance de este trabajo, podría estudiarse más afondo el tema de las covariables que varían en el tiempo, un Índice de pronóstico para los modelos no paramétricos y riesgos competitivos. Espero que los temas tratados a lo largo de este trabajo sirvan de preámbulo para su posterior estudio y desarrollo teórico.

Capítulo 5

Apéndice

5.1. Código en R

5.1.1. Base *std*

```
library(survival)
library(timereg)
library(KMsurv)
library(MASS)

data(std, package="KMsurv") #Base de datos de enfermedades de
transmision sexual.
attach(std)
```

Análisis exploratorio de datos

```
#Tiempo y censura
hist(time, freq = T, main="time")
barplot(table(rinfct), main="rinfct", names.arg = c("Censura","Reinfeccion"))

#Demograficas
hist(age, main="age")
barplot(table(race), main="race",names.arg = c("Caucasica","Negra"))
barplot(table(marst), main = "marital", names.arg = c("No casada", "casada"))
hist(yschool, main="yschool")
barplot(table(gono), main = "gono", names.arg = c("Ausente","Infectado"))
barplot(table(chla), main = "clam", names.arg = c("Ausente","Infectado"))

#Conductuales
boxplot(npartner, main="npartner")
barplot(table(os12m), main = "os12m",names.arg = c("No","SI"))
barplot(table(os30d), main = "os30d",names.arg = c("No","SI"))
barplot(table(condom),
```

```
main = "condom",names.arg = c("Siempre","Frecuente", "Nunca"))
barplot(table(rs12m), main = "rs12m",names.arg = c("No","SI"))
barplot(table(rs30d), main = "rs30d",names.arg = c("No","SI"))

#Sintomaticas
barplot(table(abdpain), main = "abdpain",names.arg = c("Ausente","Presente"))
barplot(table(discharge), main = "discharge",names.arg = c("Ausente","Presente"))
barplot(table(dysuria), main = "dysuria",names.arg = c("Ausente","Presente"))
barplot(table(itch), main = "itch",names.arg = c("Ausente","Presente"))
barplot(table(lesion), main = "lesion",names.arg = c("Ausente","Presente"))
barplot(table(rash), main = "rash",names.arg = c("Ausente","Presente"))
barplot(table(lymph), main = "lymph",names.arg = c("Ausente","Presente"))
barplot(table(vagina), main = "vagina",names.arg = c("Ausente","Presente"))
barplot(table(dchexam), main = "dchexam",names.arg = c("Ausente","Presente"))
barplot(table(abnode), main = "abnode",names.arg = c("Ausente","Presente"))
```

Modificación de la codificación y formato de las variables.

```
#Status marital
marst = rep(1,877); for(i in 1:877){if(marital[i]!="M"){marst[i]=0}}
#Gonorrea
gono = rep(0,877); for(i in 1:877){if(iinfct[i]==1|iinfct[i]==3){gono[i] = 1}}
#Clamidia
chla = rep(0,877); for(i in 1:877){if(iinfct[i]==2|iinfct[i]==3){chla[i] = 1}}
#Menor de edad
less17 = rep(0,877); for(i in 1:877){if(age[i]<17){less17[i] = 1}}
#Parejas sexuales
npar = npartner; for(i in 1:877){if(npartner[i]==0){npar[i] = 1}}
```

Estratificación de la estimación de la función de supervivencia.

```
sur =Surv(time,rinfct) #Objeto de supervivencia

#Estratificacion segun el tipo de enfermedad
plot(survfit(sur~iinfct, std), main="iinfct", conf.int = F,
mark.time = F, lty=c(1,5,3), ylab="S(t)", xlab="Dias")
legend(0,0.32, legend=c("Gonorrea", "Clamidia", "Ambas"), y.intersp = 0.8,
text.width= 150,lty=c(1,5,3))
survdiff(sur~iinfct) # Prueba de LogRank

#Estratificacion por grupos de edades
qage = c(); qage[which(age < 17)] = 1; qage[which(age > 17 & age <= 19)] = 2
qage[which(age > 19 & age <= 22)] = 3; qage[which(age > 22)] = 4
plot(survfit(sur~qage, std), main="age", conf.int = F, mark.time = F,
lty=c(1,4,3,5),ylab="S(t)", xlab="Dias")
legend(0,0.31, legend=c("age<17","17<age<19","19<age<22","22<age"),
y.intersp = 0.6, text.width= 250,lty=c(1,4,3,5))
survdiff(sur~qage) #Prueba de LogRank
```

Modelo de riesgos proporcionales paramétrico

Selección de variables.

```
full1 = coxph(sur~race+marst+less17:yschool+gono+chla+os12m+os30d+rs12m
            +rs30d+abdpain+discharge+dysuria+condom+itch+lesion+rash
            +lymph+npar:vagina+dchexam+abnode,std)
#Seleccion con el algoritmo stepAIC en reversa
stepAIC(full1, direction = "backward")

#Modelo simple obtenido
fit1 = coxph(sur~chla+os12m+condom+less17:yschool,std)
```

Evaluación de las transformaciones

```
logpart =log(npar) #Transformacion de log(npartner)

#Residuales para evaluar log(npartner)
fit.npar = coxph(sur~chla+os12m+condom+less17:yschool,std)
plot(npartner,fit.npar$residual,xlab="nparter", ylab="M(t)",cex=0.4)
lines(lowess(npartner,fit.npar$residual), lty=2)
#Residuales con la transformacion
plot(logpart,fit.npar$residual,xlab="log(npartner)", ylab="M(t)",cex=0.4)
lines(lowess(logpart,fit.npar$residual), lty=2

#Modelo definitivo
fit2 = coxph(sur~chla+os12m+condom+less17*yschool+logpart,std)
```

Búsqueda de puntos de influencia

```
#Residuales de tipo Score
res.sco = residuals(fit2,type="score",weighted=FALSE)

#Clamidia
plot(1:877,res.sco[,1],cex=0.5,xlab="No. de caso",main="chla",ylab="U(t)")
#os12m
plot(1:877,res.sco[,2],cex=0.5,xlab="No. de caso",main="os12m",ylab="U(t)")
#condom
plot(1:877,res.sco[,3],cex=0.5,xlab="No. de caso",main="condom",ylab="U(t)")
#Interaccion con less17 y yschool
plot(less17*yschool,res.sco[,4],cex=0.5,xlab="less17:yschool",
main="less17:yschool",ylab="U(t)")
#log(npartner)
plot(logpart,res.sco[,5],cex=0.5,xlab="log(npartner)",
main="log(npartner)",ylab="U(t)")
```

Residuales de devianza

```
IP=predict(fit2,type="lp") #Indice de pronostico
res.des = residuals(fit2,type="deviance") #Residuales de martingala
res.mar = residuals(fit2,type="martingale") #Residuales de devianza
plot(IP,res.mar,cex=0.4,xlab="Indice de pronostico",ylab="M(t)")
plot(IP,res.des,cex=0.4,xlab="Indice de pronostico",ylab="d")
```

Evaluacion de proporcionalidad

```
plot(survfit(sur~chla, std), lty=c(1,2),mark.time=F,fun="cloglog",xlab="Dias",
ylab="log(-log(S(t)))") # Estratificacion de clamidia
legend(100,-3,legend=c("Ausente","Presente"),text.width=0.5,lty=c(1,5))

plot(survfit(sur~os12m, std), lty=c(1,2),mark.time=F,fun="cloglog",xlab="Dias",
ylab="log(-log(S(t)))") # Estratificacion de os12m
legend(100,-3,legend=c("No","Si"),y.intersp=0.3,text.width=1,lty=c(1,5))

plot(survfit(sur~condom, std), lty=c(1,2,3), mark.time=F,fun="cloglog",xlab="Dias",
ylab="log(-log(S(t)))") # Estratificacion de condom
legend(100,-3, legend=c("Siempre","A veces","Nunca"),lty=c(1,5))

plot(survfit(sur~less17, std), lty=c(1,2), mark.time=F,fun="cloglog",xlab="Dias",
ylab="log(-log(S(t)))") # Estratificacion de less17
legend(100,-3,legend=c("No","Si"),y.intersp=0.3,text.width=1,lty=c(1,5))

# Estratificacion de npartner
fac.npar = rep(0,877); fac.npar[which(logpart>0)] = 1
plot(survfit(sur~fac.npar,std) ,lty=c(1,2),mark.time=F,fun="cloglog",xlab="Dias",
ylab="log(-log(S(t)))")
legend(100,-3,legend=c("0 parejas", ">0 parejas"),y.intersp=0.4,lty=c(1,5))

cox.zph(fit2,transform = "identity") # Prueba formal de proporcionalidad
```

Grupos de riesgo

```
q = quantile(IP, prob=c(0,0.33,0.66,1)) # Estratificacion por grupo de riesgos
riesgo = rep(2,877)
for(i in 1:877){if(IP[i]<=q[2]){riesgo[i] = 0}
  else if(IP[i]>q[2]&&IP[i]<=q[3]){riesgo[i] = 1}}

plot(survfit(sur~riesgo,std),lty=c(1),mark.time=F,xlab="Dias",
ylab="S(t)",xlim=c(0,900)) # Estimacion de KM

# Predicciones del modelo de Riesgos proporcionales
X = less17*yschool; fit=coxph(sur~chla+os12m+condom+logpart+X,std)
for(i in 0:2){
```

```
  # Vector de medias a evaluar en el modelo
  medias = data.frame(chla=mean(chla[riesgo==i]),
  os12m=mean(os12m[riesgo==i]),condom=mean(condom[riesgo==i]),
  logpart=mean(logpart[riesgo==i]),X=mean(X[riesgo==i]))
  plot(survfit(fit,newdata=medias),mark.time=F,conf.int=F,xlim=c(0,900),
  col="purple",lty=c(2),xlab="Dias",ylab="S(t)")
  par(new=TRUE)
}
legend(0,0.4, legend=c("Observados", "Predichos"),lty=c(1,2))
```

Modelo de riesgos proporcionales no paramétrico

Ajuste del modelo

```
fit.mul=timecox(sur~chla+os12m+logpart+condom+less17:yschool,std,max.time=900)
plot(fit.mul, ylab="B(t)",xlab="Dias") # Coeficientes acumulados

# Modelo con coeficientes constantes para la estimacion de sus valores
fit.mul.p = timecox(sur~const(chla)+const(os12m)+const(logpart)+
const(condom)+const(less17*yschool),data = std, max.time=900)
```

Score Process

Para mostrar las gráficas del Score process comentadas en la **sección 3.2**, se ajusta un modelo usando la función *prop()* dentro de la función *cox.aalen()*. Ya que para esta prueba no se asumen constantes los coeficientes.

```
fit.mul.p = cox.aalen(sur~prop(chla)+prop(os12m)+prop(logpart)+
prop(condom)+prop(less17*yschool),data = std, max.time=900)

# Graficas de la prueba de proporcionalidad
plot(fit.mul.p,score=T,xlab="Dias", ylab="Score␣Process")
```

Pruebas de bondad de ajuste

Se realiza el ajuste de dos modelos distintos. Uno con la transformación $\log()$ de *npartner* y otro sin ella.

```
fit.mul1 = timecox(sur~chla+os12m+logpart+condom+less17*yschool,std,
max.time=900,residuals=1) # log()
std.mul2 = timecox(sur~chla+os12m+npartner+condom+less17*yschool,std,
max.time=900,residuals=1) # Sin log()
```

Los estratos de la variable *npartner* se elaboran considerando dos niveles: $0, 1$ para indicas escasas parejas sexuales y $(0, 19]$ para indicar varias parejas sexuales.

```
Z1 =model.matrix(~-1+cut(npartner,c(0,1,19),include.lowest = T),std)
colnames(Z1) = c("Escasos", "Varios")

resids.mul1 = cum.residuals(std.mul2,std,Z1,n.sim=1000) # Residuales con log()
par(mfrow=c(1,2)); plot(resids.mul1) # Grafica de residuales
resids.mul2 = cum.residuals(fit.mul1,std,Z1,n.sim=1000) # Residuales sin log()
par(mfrow=c(1,2)); plot(resids.mul2) # Grafica de residuales
```

Para la variable *chla* se realiza el mismo análisis, suponiendo al transformación de la variable *npartner* del modelo *fit.mul1*:

```
Z = model.matrix(~-1+chla, std) # Matriz de estratificacion
resids.mul3 = cum.residuals(fit.mul1, std,Z,n.sim=1000); plot(resids.mul1)
```

Modelo aditivo

Ajuste del modelo

```
# Modelo no parametrico
fit.ad = aalen(sur~chla+os12m+condom+yschool,max.time=900,std)
par(mfrow=c(2,3));plot(fit.ad, xlab="Dias", ylab="B(t)") # Coeficientes
```

Prueba de hipótesis para efecto invariante

Para probar la hipótesis $H_0 : \beta_j(t) = \gamma$ mediante el *Test score*, se debe justar un modelo semi-paramétrico en el que se suponga que los coeficientes a probar no son constantes, en este caso *condom*. Eso se obtiene mediante el siguiente comando:

```
fit.semi.ad = aalen(sur~const(chla)+const(os12m)+condom+const(yschool),
data = std, max.time=900)

# Grafica del test process
par(mfrow=c(1,2)); plot(fit.semi.ad,score=T,xlab="Dias",ylab="Test␣Process")

# Modelo parametrico
fit.ad.p = aalen(sur~const(chla)+const(os12m)+const(condom)+const(yschool),
data = std, max.time=900)
```

Bondad de ajuste

```
fit.res = aalen(sur~chla+os12m+condom+yschool,max.time=900,std,residuals=1,
n.sim=0)# Modelo con residuales
X=model.matrix(~-1+cut(condom,c(0,1,2,3),include.lowest = T),std);
colnames(X)=c("Siempre","A␣veces","Nunca");
resids=cum.residuals(fit.res,std,X,n.sim=1000)
par(mfrow=c(2,2)); plot(resids, hw.ci = 2); summary(resids)
```

Grupos de riesgo

Como en la sección anterior, se procede a elaborar los grupos de riesgo con base a los cuantiles del vector el índice de pronóstico. Elaborado con el siguiente código:

```
rel=as.matrix(data.frame(chla,os12m,condom,yschool))# Valores de las covariables del model
IP.ad=c(fit.ad.p$gamma) %*% t(rel)#Indice de pronostico

q = quantile(IP.ad, prob=c(0,0.33,0.66,1))
riesgo.ad = rep(2,877)
```

```
for(i in 1:877){if(IP.ad[i]<=q[2]){riesgo.ad[i] = 0}
  else if(IP.ad[i]>q[2] && IP.ad[i]<=q[3]){riesgo.ad[i] = 1}}
```

La estimación de la función de supervivencia estratificada se muestra a continuación:

```
for(i in 0:2){ # Predicciones
  risk.gr = rel[which(riesgo.ad==i),] # Grupos de riesgo
  z1=mean(risk.gr[,1]);z2=mean(risk.gr[,2]);z3=mean(risk.gr[,3])
  z4=mean(risk.gr[,4]); Z = c(z1,z2,z3,z4) # Valores a evaluar
  S0 = exp(-t(fit.ad.p$cum[,-1])- fit.ad.p$cum[,1]*sum(Z%*%fit.ad.p$gamma))
  plot(fit.ad.p$cum[,1],S0,lty=c(2),type="l",col="purple", xlab = "Dias",
       ylab = "S(t)",xlim=c(34,867), ylim=c(0,1))
  par(new=TRUE)
}
# Observaciones
plot(survfit(sur~riesgo.ad), mark.time = F,xlim=c(0,900))
legend(0,0.3,legend=c("Observados","Predichos"),lty=c(1,2))
```

Modelo aditivo-multiplicativo

Ajuste del modelo

Se ajustan dos modelo. Uno bajo la transformación *log*() para *npartner* y otro sin ella.

```
fit.am1 = cox.aalen(sur~prop(yschool*less17)+prop(os12m)+prop(condom)+chla+
prop(logpart),std,max.time=900,residuals=1) # Con log()
fit.am2 = cox.aalen(sur~prop(yschool*less17)+prop(os12m)+prop(condom)+chla+
prop(npartner),std,max.time=900,residuals=1) # Sin log()

par(mfrow=c(1,2)); plot(fit.am1,xlab="Dias",ylab = "B(t)") # Coeficientes
```

Bondad de Ajuste

Se crea una matriz para señalar si hay escasas o varias parejas sexuales. Tal y como se realiza en la prueba bondad de ajuste del modelo de riesgos proporcionales

```
Y2 =model.matrix(~-1+cut(npartner,c(0,1,19),include.lowest = T),std)
colnames(Y2) = c("Escasos", "Varios")
resids.admu1 = cum.residuals(fit.am1,std,Y2,n.sim=0) # Residuales de fit.am1
resids.admu2 = cum.residuals(fit.am2,std,Y2,n.sim=0) # Residuales de fit.am2
par(mfrow=c(1,2)); # Graficas de residuales
plot(resids.admu2,score=T,xlab="Dias");plot(resids.admu2,score=T,xlab="Dias")
```

Score Process

```
par(mfrow=c(2,2));plot(fit.am1,score=T, xlab="Dias", ylab= "Score␣test")
```

Grupos de riesgo

Se dividen los grupos de riesgo dentro de las dos estratificaciones definidas por la variable dicotómica *chla*.

```
plot(survfit(sur~1),mark.time=F,xlim=c(-34,925),conf.int=F)
for(j in 0:1){
  z1=mean(yschool*less17);z2=mean(os12m);z3=mean(npar);
  z4=mean(condom);x0 = c(1,j) # Evaluaciones en la media
  Z=c(z1,z2,z3,z4); # Valores de evaluacion en el modelo
  #Funcion de supervivencia
  S0 = c(exp(-exp(sum(Z*fit.am1$gamma))*(x0 %*% t(fit.am1$cum[,-1]))))
  par(new=TRUE)
  plot(fit.am1$cum[,1],S0,col=c("darkblue"),type="s",ylim=c(0,1))
  legend(0,0.18,legend=c("Observados","Predichos"),lty=c(1,2))
  }
```

5.1.2. Base *pbc*

Modelo de riesgos proporcionales paramétrico

```
library(randomForestSRC)
data(pbc, package="randomForestSRC") # Enfermedades de transmision sexual.
```

Análisis exploratorio de datos

```
pbc312 = pbc[1:312,] # Pacientes con DPCA administrado aleatoriamente
attach(pbc312)
sur = Surv(days/360,status)
plot(survfit(sur~treatment,pbc),conf.int=F,mark.time=F) #Estratos de tratamiento

bili.cuar = rep(4,418) #Cuartiles de bili
bili.cuar[bili>0&bili<=0.8] = 1; bili.cuar[bili>0.8&bili<=1.3] = 2;
bili.cuar[bili>1.3&bili<=3.4] = 3

# Estimador Kaplan-Meier para cada estrato de cuartil de bili
plot(survfit(sur~bili.cuar,pbc),conf.int=F,mark.time=F,lty=c(1,2,3,4))
```

Selección de variables

Ajuste de cada uno de los modelos ajustados por *Fleming & Harrington* (1991, pág. 159, tabla 4.4.2 y 4.4.3)

```
pbc.full1 = coxph(sur~age+albumin+alk+ascites+bili+edema+hepatom+
platelet+prothrombin+sex+spiders,pbc312)
pbc.short1 = coxph(sur~age+albumin+bili+edema+hepatom+prothrombin)
pbc.full2 = coxph(sur~age+log(age)+albumin+log(albumin)+bili+
log(bili)+edema+prothrombin+log(prothrombin)+hepatom,pbc312)
pbc.mid2 = coxph(sur~age+albumin+log(bili)+edema+prothrombin,pbc312)

# Modelos sobre el que se hace el resto del analisis
pbc.main=coxph(sur~age+log(albumin)+log(bili)+edema+log(prothrombin),pbc312)
```

Evaluación de las transformaciones

Residuales de martingalas para el modelo con las variables *age*, *edema*, log(*albumin*) y log(*prothrombin*) vs. *bili* para distintos grupos de observaciones de la base *pbc*

```
# 312 pacientes no aleatorios
pbc.312 = coxph(sur~age+log(albumin)+log(prothrombin)+edema,pbc312)
res1 = pbc.312$residuals
plot(log(bili),res1, xlab = "Log(bili)", ylab="M(t)")
lines(lowess(log(bili),res1), lty=2) # Con log()
plot(bili,res1,xlab = "bili", ylab="M(t)")
lines(lowess(bili,res1), lty=2) # Sin log()

h.152=which(hepatom==0) #152 sin hepatomegalia
pbc.152 = coxph(Surv(days[h.152]/360,status[h.152])~age[h.152]+
log(albumin[h.152])+log(prothrombin[h.152])+edema[h.152])
res2 = pbc.152$residuals # Residuales de martingala
plot(log(bili[h.152]),res2,xlab="log(bili)",ylab="M(t)") # Con log()
lines(lowess(log(bili[h.152]),res2), lty=2)
plot(bili[h.152],res2, xlab = "bili", ylab = "M(t)") # Sin log()
lines(lowess(bili152,res2), lty=2)

h.160=which(hepatom==1) #160 con hepatomegalia
pbc.160 = coxph(Surv(days[h.160]/360,status[h.160])~age[h.160]+
log(albumin[h.160])+log(prothrombin[h.160])+edema[h.160])
res3 = pbc.160$residuals # Residuales de martingala
# Con log()
plot(log(bili[h.160]),res3,xlab="log(bili)", ylab="M(t)")
lines(lowess(log(bili[h.160]),res3), lty=2)
# Sin log()
plot(bili[h.160],res3,xlab="bili",ylab="M(t)")
lines(lowess(bili[h.160],res3), lty=2)
```

Búsqueda de puntos de influencia

Residuales de score de cada variable vs. cada variable (o el número de caso para los residuales de *edema*).

```
res.sco = residuals(pbc.main, type = "score", weighted = FALSE)

plot(age, res.sco[,1], xlab = "age", ylab = "U(t)")
plot(log(albumin),res.sco[,2], xlab = "log(albumin)",ylab="U(t)")
plot(log(bili), res.sco[,3], xlab = "log(bili)", ylab = "U(t)")
plot(1:312, res.sco[,4], xlab="edema",ylab="U(t)")
plot(log(prothrombin), res.sco[,5], xlab="log(prothrombin)",ylab="U(t)")
```

Residuales de devianza

Se comparan los residuales de devianza con los residuales de martingala. Ambos se grafican vs. el índice de pronóstico.

```
# Indice de pronstico
risk.score = as.matrix(data.frame(age, log(albumin),log(bili),edema,
log(prothrombin))) %*% pbc.main$coefficients
plot(risk.score, pbc.main$residuals, cex=0.5,xlab= "Indice", ylab= "M(t)")

#Transformacion de devianza
dev = residuals(pbc.main, type="deviance")
plot(risk.score,dev, cex = 0.5, ,xlab= "Indice de pronostico",ylab="d(t)")
```

Evaluación de la proporcionalidad

Se elaboran estratos con los cuartiles de cada variable (o niveles para las variables discretas) y se grafican el estimador de Kaplan-Meier bajo la transformación $log(-logS(t))$.

```
age.fact = rep(4,312); age.fact[age>61] = 1 # Estratos de Age
age.fact[age>53&age<=61] = 2; age.fact[age>46&age<=53] = 3

plot(survfit(sur~age.fact,pbc312), mark.time=F,fun = "cloglog")

# Estratos de Edema
plot(survfit(sur~edema,pbc312),mark.time=F,fun="cloglog")

pro.fact = rep(4,312); pro.fact[prothrombin>11.18] = 1
pro.fact[prothrombin>11&prothrombin<=11.18] = 2
pro.fact[prothrombin>10.5&prothrombin<=11] = 3 # Estratos de Prothrombin

plot(survfit(sur~pro.fact,pbc312),mark.time=F,
fun = "cloglog", ylim=c(-6,2))

# Estratos de Bili
plot(survfit(sur~bili.cuar,pbc312),mark.time=F,fun = "cloglog", ylim=c(-6,2))

# Estratos de Albumin
alb.fact = rep(0,312); alb.fact[albumin>3.67] = 1
alb.fact[albumin>3.4&albumin<=3.67] = 2
alb.fact[albumin>3.1&albumin<=3.4] = 3

plot(survfit(sur~alb.fact,pbc312),mark.time=F,fun = "cloglog")
```

Función de supervivencia estimada

```
q = quantile(risk.score, prob=c(0,0.33,0.66,1))
riesgo = rep(2,312)# Estratificacion por grupo de riesgos
for(i in 1:312){if(risk.score[i]<=q[2]){riesgo[i] = 0}
  else if(risk.score[i]>q[2]&&risk.score[i]<=q[3]){riesgo[i] = 1}}

plot(survfit(sur~riesgo,pbc312),lty=c(1)ylab="S(t)") # Estimacion de KM

# Especificacion del modelo para estimar
logAlbumin=log(albumin); logBili=log(bili); logPro=log(prothrombin)
fit=coxph(sur~age+logAlbumin+logBili+edema+logPro,pbc312)
for(i in 0:2){
  # Vector de medias a evaluar en el modelo
  medias=data.frame(age=mean(age[riesgo==i]),
  logAlbumin=mean(log(albumin[riesgo==i])),logBili=mean(log(bili[riesgo==i])),
  edema=mean(edema[riesgo==i]),logPro=mean(log(prothrombin[riesgo==i])))
  plot(survfit(fit,newdata=medias),conf.int=F,col="purple"); par(new=TRUE)
}
legend(0,0.27,legend=c("Observados","Predichos"),col=c("black","purple"))
```

Modelo de riesgos proporcionales no paramétrico

La función *timecox()* permite, por defecto, estimar los coeficientes $\hat{B}(t)$ asumiendo que el efecto de las covariables varía a lo largo del tiempo. La función *const()* permite especificar que se asume que una covariable tiene efecto constante.

```
pbc.p = timecox(sur~age+edema+log(bili)+log(albumin)+log(prothrombin),pbc)
par(mfrow=c(2,3)); plot(pbc.cox)
```

Ajuste del modelo semiparamétrico

```
pbc.sp = timecox(sur~const(age)+edema+const(log(bili))+
const(log(albumin))+log(prothrombin),pbc)
```

Prueba de proporcionalidad

```
pbc.p2 = cox.aalen(sur~prop(age)+prop(edema)+prop(log(bili))+
prop(log(albumin))+prop(log(prothrombin)),pbc)
plot(pbc.2,score=T)
```

Prueba de bondad de ajuste

Para calcular un modelo con residuales de martingalas acumulativos, hay que indicarlo dentro de la función *timecox()* con el parámetro *residuals*. Estos se calcularán mediante la función *cum.residuals()*. Se presentan los residuales para un modelo con y sin la transformación *log()* de la variable *bili*.

```
pbc.cox3 = timecox(sur~age+edema+log(bili)+log(albumin)+log(prothrombin),pbc,
residuals=1) # con log()
pbc.cox4 = timecox(sur~age+edema+bili+log(albumin)+log(prothrombin),pbc,
residuals=1) # sin log()

#Evaluacion de log(bili)
Z =model.matrix(~-1+cut(bili,quantile(bili),include.lowest = T),pbc)
colnames(Z) = c("1. Cuantil","2. Cuantil", "3. Cuantil", "4. Cuantil")
resids.mul1 = cum.residuals(pbc.cox3,pbc,Z) # Residuales con log()
par(mfrow=c(2,2)); plot(resids.mul1) # Grafica de residuales
resids.mul2 = cum.residuals(pbc.cox4,pbc,Z) # Residuales sin log()
par(mfrow=c(2,2)); plot(resids.mul2) # Grafica de residuales
```

Modelo aditivo

Ajuste del modelo

La función **aalen()** permite ajustar el modelo de riesgos aditivo. Aquellas variables cuyo efecto se suponga constante, se especifican mediante la función **const()**. También se encuentra dentro del paquete *timereg*.

```
# Estimacion de los coeficientes
pbc.a1 = aalen(sur~age+edema+bili+log(albumin)+log(prothrombin),pbc)
summary(pbc.a1) # Prueba de significancia e invarianza de los coeficientes

par(mfrow=c(3,2)); plot(pbc.a1,ylab="B(t)") # Coeficientes de regresion

pbc.sa = aalen(sur~const(age)+edema+const(log(albumin))+const(bili)+
log(prothrombin),pbc) # Modelos semiparametrico
summary(pbc.sa) # Coeficientes estimados
```

Test process

Se puede probar la hipótesis de coeficientes $\beta(t)$ constantes para aquellas variables que pertenecen a la parte no paramétrica del modelo semi-paramétrico, usando el mismo modelo ajustado.

```
plot(pbc.sa,score=TRUE,ylab="Test process")
```

Estimacion de la función de supervivencia

```
#Puntos de evaluacion
x0 = c(1,mean(edema),mean(log(prothrombin),na.rm=T))
z0 = c(mean(age),mean(log(albumin)),mean(log(bili)))
# Funcion de supervivencia
S0 = exp(-x0 %*% t(pbc.sa$cum[,-1])-pbc.sa$cum[,1]*sum(z0*pbc.sa$gamma))
plot(pbc.sa$cum[,1],S0,type="s",ylab="S(t)",xlim=c(0,12))
par(new=TRUE)
plot(A,mark.time=F,conf.int=F,xlim=c(-0.5,12.46))
legend(0.2,0.6,legend=c("Observados","Predichos"),col=c("black","blue"))
```

Bondad de ajuste

Para el cálculo de los residuales de este modelo, se utiliza la matriz $\boldsymbol{K}(t)$ para especificar el grupo de datos con los que se hará la estimación de los residuales que se usó en el modelos de riesgos proporcionales, pues se hace el análisis con los cuartiles de la misma variable. Se ajustan modelos mediante la función *aalen()*, en la que se indica la estimación de los residuales mediante el parámetro *residuals=1* .

```
pbc.a2 = aalen(sur~age+edema+bili+log(albumin)+log(prothrombin),
pbc, residuals=1) # Modelos sin log()

resids.ad1 = cum.residuals(pbc.a2,pbc,Z) # Residuales
par(mfrow=c(2,2)); plot(resids.ad1); summary(resids.ad1) # sin log()

pbc.a3 = aalen(sur~age+edema+log(bili)+log(albumin)+log(prothrombin),
pbc, residuals=1) # Modelo con log()

resids.ad2 = cum.residuals(pbc.a3,pbc,Z1) # Residuales
plot(resids.ad2); summary(resids.ad2) # con log()
```

Modelo aditivo-multiplicativo

El ajuste se realiza mediante la función *cox.aalen()* del paquete *timereg*, donde se especifican las variables que se incorporan al modelo de manera proporcional mediante la función *prop()*, mientras que las restantes son aditivas.

```
pbc.am1 = cox.aalen(sur~prop(age)+edema+prop(log(bili))+prop(log(albumin))+
log(prothrombin),pbc) # Ajuste
summary(pbc.am1) # Pruebas de significancia y variacion de los coeficientes
```

Bondad de ajuste

```
# Con log()
pbc.am2 = cox.aalen(sur~prop(age)+edema+prop(log(bili))+prop(log(albumin))+
log(prothrombin),pbc, residuals=1)
resids.am1 = cum.residuals(pbc.am2,pbc,Z)
par(mfrow=c(2,2)); plot(resids.am1); summary(resids.am1)

# Sin log()
pbc.am3 = cox.aalen(sur~prop(age)+edema+prop(bili)+prop(log(albumin))+
log(prothrombin),pbc, residuals=1)
resids.am2 = cum.residuals(pbc.am3,pbc,Z)
par(mfrow=c(2,2)); plot(resids.am2); summary(resids.am2)
```

Estimación de la función de supervivencia

```
# Puntos de evaluacion
x0 = c(1,mean(Edema),mean(log(prothrombin),na.rm=T))
z0 = c(mean(age),mean(log(bili)),mean(log(albumin)))
```

```
# Funcion de supervivencia
S0 = c(exp(-exp(sum(z0*pbc.am1$gamma))*(x0 %*% t(pbc.am1$cum[,-1]))))
plot(pbc.am1$cum[,1],S0,type="s",ylab="S(t)",col="red",lty=2)
par(new=TRUE)
}
plot(survfit(sur~1,pbc),mark.time=F,conf.int=F,ylab="S(t)")
legend(0.2,0.4,legend=c("Observados","Predichos"),lty=1:2)
```

Bibliografía

- AALEN O. O., BORGAN Ø. & H. K. GJESSING (2008), *Survival and event history analysis: A process point of view*, Statistics for biology and health, Springer Science+Business Media, New York.
- ANDERSEN P., BORGAN Ø., GILL R. & KEIDING N. (1993), *Statistical models based on counting processes, Springer verlag, New York.*
- COX D. R & OAKES D. (1984), *Analysis of survival data*, Monographs on statistics and applied probability, Capman & Hall, New York.
- DABROWSKA D.M. (1997), *Smoothed Cox regression*, Ann. Statistic 25, pág. 1510-1540.
- FLEMING T. R. & HARRINGTON D. P. (1991), *Counting process and survival analysis*, Wiley series in probability and mathematical statistics, United states of America.
- KLEIN J. P. & MOESCHBERGER M. L. (2003), *Survival Analysis: Techniques for censored and truncated data*, Statistics for biology and health, Springer-Verlag, New york.
- LAWLESS J. F. (2003), *Statistical Models and Methods for lifetime data*, Wiley series in probability and statistics, John Wiley & Sons, New Jersey.
- McCULLAGH P. & NELDER J. A. (1989) *Generalized linear models*, Chapman & Hall, London.
- MARSAGLIA G. & TSANG W. W. (2003). *Evaluating Kolmogorov's Distribution*. Journal of Statistical Software. 8 (18): 1–4.
- MARTINUSSEN T. & SCHEIKE T. (2006), *Dynamic regression models for survival data*, Springer Science + Business Media, New York.
- PARMAR M. & MACHIN D. (1995), *Survival analysis: a practical approach*, John Wiley & Sons, Cambridge.
- THERNEAU T. M. & GRAMBSCH P. M. (2000), *Modeling survival data: Extending de Cox model*, Statistics for biology and health, Springer Science+Business Media, New York.