**UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS**

FACULTAD DE CIENCIAS MATEMÁTICAS

E.A.P. DE ESTADÍSTICA

# Modelo de regresión de Cox usando splines

TESIS

Para optar el Título Profesional de Licenciado en Estadística

AUTOR

Claudio Jaime Flores Flores

ASESOR

Antonio Bravo Quiroz

**Lima - Perú**

**2011**

MODELO DE REGRESIÓN DE COX USANDO SPLINES

CLAUDIO JAIME FLORES FLORES

Tesis presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el Título Profesional de Licenciado en Estadística.

Aprobado por:

---

Mg. Violeta Alicia Nolberto Sifuentes
(Presidente)

---

Mg. María Estela Ponce Aruneri
(Miembro)

---

Mg. Antonio Bravo Quiroz
(Miembro Asesor)

LIMA - PERU
Octubre - 2011

**F**ICHA CATALOGRAFICA

FLORES FLORES, CLAUDIO JAIME

– – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – –

Modelo de Regresión de Cox usando Splines.
ix, 97p., 29.7 cm, (UNMSM, Licenciado, Estadística, 2011).
Tesis Profesional, Universidad Nacional Mayor de San Marcos,
Facultad de Ciencias Matemáticas 1. Estadística I UNMSM
FdeCM. Título (serie).

DEDICATORIA:

A mis padres: Alejandro (en memoría) y Brígida, y mis hermanos por todo el apoyo durante mi formación.

A mi familia: mi esposa Liliana y mi hija María José, por todo el apoyo y su comprensión para hacer realidad este trabajo.

AGRADECIMIENTOS:

A los profesores de la Escuela Académica Profesional de Estadística - UNMSM, en particular a los profesores Fátima Medina, Violeta Nolberto, Antonio Bravo e Ysela Agüero, por todo el apoyo durante mi formación en Pre-Grado.

A los Profesores Guadalupe Gómez y Pedro Delicado de la Universidad Politécnica de Cataluña por su apoyo durante mi formación académica (España).

# Índice

# RESUMEN


MODELO DE REGRESION DE COX USANDO SPLINES

CLAUDIO JAIME FLORES FLORES

Octubre - 2011

Asesor: Mg. Antonio Bravo Quiroz
Título obtenido: Ingeniero Estadístico

---

En muchos estudios clínicos es muy frecuente el uso de modelo de riesgos proporcionales de Cox; el cual asume riesgos proporcionales y restringe a que el logaritmo de la razón de riesgo sea lineal en las covariables, lo cual en muchos casos no se verifica. En este sentido, una forma funcional no lineal del efecto de las covariables puede ser aproximada por una función spline. En este trabajo, se presenta la metodología del modelo de regresión de Cox usando splines, particularmente regresión splines y P-splines, para aproximar la forma funcional no-lineal de los efectos de las covariables en la función de riesgo. Como una aplicación, se analiza los datos de pacientes con LNH para determinar los factores pronósticos para la supervivencia global. Los resultados muestran que el efecto de las covariables contínuas como hemoglobina, leucocitos, linfocitos y DHL presentan una forma funcional no lineal en el logaritmo de la razón de riesgo.



**Palabras claves:** Modelo de Cox, regresión splines, P-splines, LNH.

# ABSTRACT

**C**OX REGRESSION MODEL USING SPLINES

CLAUDIO JAIME FLORES FLORES

October - 2011

Advisor: Mg. Antonio Bravo Quiroz
Title obtained: Ingeniero Estadístico

– – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – – –

In many clinical studies, Cox proportional hazard model is very common to use, it assumes proportional hazard and restricts the log hazard ratio to be linear in the covariates; these asumptions can not be verified. In this way, a nonlinear functional form of the covariates effect can be approximated by a spline function. In this paper, we present the methodology and an application of Cox model using splines, particularly regression splines and P-splines, to approximate the nonlinear functional form of the effect of covariates on the hazard function. As an application, we analyse data from patients with NHL to determine prognostic factors for overall survival. These results show that the effect of continuous covariates as: hemoglobin, leukocytes, lymphocytes and LDH have a nonlinear form with the log hazard ratio.

**Keywords:** Cox model, regression splines, P-spline, NHL.

# Capítulo 1
# Introducción

## 1.1. Introducción

Una particularidad de las técnicas estadísticas que se plantean para analizar un conjunto de datos observacionales o experimentales, es cuando la variable de estudio corresponde al tiempo de seguimiento hasta la ocurrencia de un evento de interés (muerte, recurrencia, progresión, complicación, etc.).

En muchos estudios clínicos el tiempo de seguimiento hasta la ocurrencia de un evento de interés, puede ser el tiempo de supervivencia de un grupo de pacientes sometidos a un tipo de tratamiento. Estos datos son usualmente conocidos como datos de supervivencia, y en general las muestras son incompletas o censuradas, por lo que las técnicas estadísticas relacionadas a este tipo de datos se denominan análisis de supervivencia.

El análisis de supervivencia es una área de la estadística que comprende un conjunto de técnicas y modelos, para analizar el tiempo que transcurre entre un evento inicial (fecha de diagnóstico, fecha de tratamiento, etc) y un evento final (evento de interés), llamado tiempo de supervivencia, y en algunos casos en presencia de variables explicativas denominadas covariables. Dichos análisis son, hoy en día, una parte fundamental en muchos estudios clínicos, ensayos clínicos, estudios epidemiológicos y de muchas otras disciplinas como la economía, la ciencia actuarial y la ingeniería.

En muchos estudios clínicos es muy común que además de ser observado el tiempo de supervivencia sean observadas las características clínicas, llamadas covariables. Si el interés es en determinar el efecto de las covariables en la supervivencia, el análisis se reduce en realizar análisis de regresión. En análisis de supervivencia, el modelo de regresión frecuentemente utilizado para analizar el efecto de las covariables en la supervivencia, es el modelo de riesgos proporcionales de Cox (1972).

En el modelo de Cox la respuesta modelada es la función de riesgo, con logaritmo de la razón de riesgo que depende de las covariables en una forma lineal; esto implica, que la

razón de riesgo no varía en el tiempo, el riesgo para dos individuos es proporcional y el efecto de las covariables presenta una relación lineal. Sin embargo, estas suposiciones son muy restrictivas y en muchos casos pueden no verificarse, y en consecuencia las estimaciones bajo el modelo de Cox clásico serían incorrectas.

En situaciones de no cumplimiento del supuesto de riesgos proporcionales, el modelo de Cox estratificado, el modelo de Cox con variables tiempo-dependientes, el modelo de Cox ponderado, el modelo de Odds proporcional o el modelo log-logístico podrían ser una buena alternativa. Sin embargo, en situaciones en que el efecto de las covariables no presenten una relación de forma funcional lineal, estos modelos no serían los más adecuados para analizar los datos.

Durante las últimas décadas numerosas técnicas se han desarrollado para aproximar la forma funcional del efecto de las covariables de una manera más flexible, utilizando para ello métodos de suavizamiento con splines.

En este trabajo se presenta la metodología del modelo regresión de Cox con splines para aproximar la forma funcional no lineal del efecto de las covariables contínuas y una aplicación para determinar los factores pronósticos para la supervivencia global de los pacientes con linfoma no Hodgkin.

## 1.2. Objetivos del trabajo

El objetivo de este trabajo consiste en presentar los aspectos metodológicos del modelo de regresiópn de Cox con splines, particularmente modelo de Cox con regresión splines y modelo de Cox con P-splines, para aproximar la forma funcional no lineal del efecto de las covariables en la función de riesgo. Asi mismo la aplicación de esta metodología para determinar los factores pronósticos para la supervivencia de pacientes con Linfoma no Hodgkin (LNH) diagnósticados y tratados en el Instituto Nacional de Enfermedades Neoplásicas (INEN) entre 1990 a 2002.

Los objetivos del específicos son dos:

-Describir los aspectos metodológicos del modelo de Cox con regresión splines y modelo de Cox con P-splines para aproximar la forma funcional no lineal del efecto de la covariables contínuas en la función de riesgo.

-Determinar los factores pronósticos para la supervivencia en pacientes con LNH y determinar la forma funcional del efecto de las covariables en la razón de riesgo (hazard ratio), que no pueden ser vistas cuando se utilizan procedimientos clásicos, mucho menos con variables continuas categorizadas.

## 1.3. Detalles del desarrollo

El contenido de este trabajo se encuentra estructurado en 6 capítulos. En el capítulo 2 se presenta el aspecto clínico y pronóstico de los LNH. En el capítulo 3 se presentan los conceptos básicos relacionados al análisis de supervivencia. En la subsección 3.1 se hace una descripción de los datos utilizados en análisis de supervivencia, en la subsección 3.2 las funciónes de distribución, en la sección 3.3 la función de máxima verosimilitud, en la subsección 3.4 el procesos de conteo y en la subsección 3.5 el método de suavizamiento splines.

En el capítulo 4 se describe la base teórica del modelo de regresión de Cox usando splines, particularmente del modelo de Cox con regresión splines y modelo de Cox con P-splines. En la subsección se 4.1 se hace una revisión de literatura, en la subsección 4.2 se describe brevemente el modelo de Cox clásico, en la subsección 4.3 se hace una breve introducción al modelo de Cox con splines, en la subsección 4.4 se describe el modelo de Cox con regresión splines, en la sección 4.5 se describe el modelo de Cox con P-spliens y en la sección 4.6 se describe los métodos de diagnóstico del modelo de Cox.

En el capítulo 5 se realizan las aplicaciones de los modelos descritos en la sección 4 (modelo de Cox clásico, modelo de Cox con regresión splines y modelo de Cox con P-splines) para determinar los factores pronósticos y la forma funcional de los efectos de las covariables en la supervivencia global de los pacientes con LNH.

En el capítulo 6 se da una breve discusión y las conclusiones de este trabajo, así como las recomendaciones para los trabajos futuros.

# Capítulo 2
# LNH: Aspectos clínicos y pronósticos

En este capítulo se presenta los aspectos clínicos del linfoma no Hodgkin, de manera que, nos permita conocer un poco sobre la naturaleza de la enfermedad y su pronóstico. Se describe la enfermedad en su aspecto clínico epidemiológico, la supervivencia y los factores pronósticos para la supervivencia según la literatura especializada. Así mismo, se dá una breve descripción de los datos que son objeto de análisis para determinar los factores pronósticos para la supervivencia global utilizando métodos de suavizamiento con splines que son descritos en el capítulo 3.

## 2.1. Linfoma No Hodgkin

El cáncer es uno de los principales problemas de salud pública en el mundo y ocupa el segundo lugar entre las causas de muerte después de las enfermedades cardiovasculares. Por otro lado, debido a que el cáncer es una enfermedad potencialmente curable en etapas tempranas, es importante disponer de indicadores que permitan un mejor seguimiento en términos de incidencia, mortalidad y supervivencia (Programas Nacionales de Control de Cáncer, OPS 2004).

Los linfomas no Hodgkin (LNH) son neoplasias linfoproliferativas del sistema linfático y constituyen un grupo muy heterogéneo de enfermedades definidas por aspectos morfológicos, inmunofenotipos y genéticos. Cuando las células linfáticas mutan y se proliferan sin estar reguladas por los procesos que habitualmente controlan el crecimiento y la muerte celular, se forman tumores en las áreas donde existe el tejido linfático y pueden diseminarse a cualquier órgano (Friedberg y cols, 2008).

La etiología del LNH se desconoce en la mayoría de los casos, sin embargo, existen situaciones clínicas en que se presenta una mayor incidencia de procesos linfoproliferativos debido a estados de inmunodeficiencia y trastornos en proceso de inmunorregulación. Las causas asociadas a las inmunodeficiencias adquiridas pueden ser debido a infecciones por el virus

de Epstein Barr (EBV), virus linfotrópico humano tipo 1 (HTLV-1), virus de la inmunodeficiencia humana (HIV), virus de la hepatitis C (HCV), herpes virus humano 8 (HHV-8), Helicobacter Pylori y por exposiciones a radiaciones, fármacos entre otros (Friedberg y cols (2008), Hartge (2007)).

El LNH representa la décima neoplasia más frecuente en el mundo y su incidencia varía entre los diferentes países, regiones del mundo y periodos de estudio según la Agencia Internacional para la Investigación en Cáncer (IARC). Las tasas más elevadas se observan en los países más desarrollados como Norteamérica, Europa Occidental, Oceanía y las más bajas en India y los países Africanos. A nivel mundial se estimaron para el año 2000 alrededor de 10,1 millones de nuevos casos y 6,2 millones de muertes por cáncer; de los cuales, el LNH representa 2.9 % de nuevos casos y 2.6 % de muertes por cáncer, después del cáncer de pulmón, mama, colorectal y estómago (Muir y cols (1987), Parkin y cols (2002)).

En el Perú, para el año 2000 se diagnosticaron 1767 nuevos casos de LNH, que representa la quinta neoplasia mas frecuente después de cáncer de estómago, cérvix, próstata, mama, colorectal, pulmón y hígado, con una incidencia de 8.3 x 100.000 personas, tercer país con mayor incidencia en Sudamérica después de Uruguay y Bolivia, aunque primero en las mujeres (Globocan 2002).

El tratamiento de los pacientes con LNH, depende del grado de agresividad (indolente y agresivo) y el estadio clínico o el índice pronóstico internacional (IPI) que clasifica a los pacientes en cuatro grupos de riesgo (bajo, intermedio bajo, intermedio alto y alto) teniendo en cuenta la edad, estado funcional, estadio clínico, deshidrogenasa láctica y ganglios extraganglionares. La modalidad de tratamiento puede ser radioterapia (Rt), quimioterapia (Qt) e inmunoterapia; aunque en la mayoría de los casos se utiliza la quimioterapia como la forma principal de tratamiento. El esquema CHOP (ciclofosfamida, doxorubicina, vincristina y prednisona) se considera aún como un régimen de quimioterapia estándar en el tratamiento de este grupo de pacientes (Arece y Rodríguez (2003), Friedberg y cols (2008)).

## 2.2. Supervivencia y pronóstico

La tasa de supervivencia a 5 años de los pacientes con LNH varía aproximadamente entre 50 % y 70 %. Los linfomas indolentes tienen un pronóstico relativamente bueno, con mediana de supervivencia de hasta 10 años, pero generalmente no son curables en estadios clínicos avanzados (EC III-IV). En cambio los linfomas agresivos tienen una historia clínica natural más corta, pero un número significativo (entre 30 % y 60 %) de estos pacientes pueden curarse con regímenes agresivos de quimioterapia en combinación (Kyle y Hill (2010)).

Según la literatura, el pronóstico de los pacientes con LNH depende de múltiples factores; siendo los más relevantes: la edad, estado de performance, estadio clínico, deshidrogensa láctica (DHL), $\beta$2-microglobulinas ($\beta 2M$), tipo histológico, tipo celular (linaje B o T) y grado de agresividad. En algunas series publicadas, los pacientes mayores de 60 años de edad, con enfermedad ganglionar, escala ECOG 2-4, estadio clínico III-IV, síntomas B, DHL elevada ($>240U/L$), $\beta 2M$ elevada ($>3.5$), linfoma agresivo, linfoma de células T, pacientes con hemoglobina baja ($<12g/dl$), leucocitos elevados ($>10mil$) y linfocitos elevados ($>40\,\%$) presentan una pobre tasa de supervivencia a 5 años (Mounier, et.al (1997), Horsman y Hancock (2001), Rabasa (2001)).

## 2.3. Factores pronósticos

Desde la década de los años 70 ha existido un enorme interés en la comunidad científica por el estudio de los factores pronósticos (FP) en los linfomas, como prototipo de enfermedad curable. Probablemente los linfomas sean las neoplasias mejor y más ampliamente estudiadas; sin embargo, se precisan de nuevos estudios para clarificar su utilidad, debido a que un número relativamente importante de pacientes presentan recurrencia o que fallecen a consecuencia de la enfermedad, la aparición de nuevos FP (hemoglobina, leucocitos, linfocitos y marcadores tumorales) y el desarrollo de nuevos métodos estadísticos para su análisis.

Según la literatura, los factores pronósticos en los pacientes con LNH se agrupan en tres grandes grupos, aquellos que se derivan de las características del paciente, del tumor y del tratamiento. Dentro de los FP dependientes del tumor, se tiene en cuenta las características biológicas y la carga tumoral (Mounier, et al. (1997), Costas, et al. (1998), Horsman y Hancock (2001), Rabasa 2002)).

- Dentro de los factores pronósticos dependientes del paciente se considera la edad, el estado funcional, las enfermedades preexistentes y la competencia inmunológica. La edad se considera como un factor pronóstico, debido a que esta se asocia a una mayor morbimortalidad después de los 60 años. El estado funcional según la escala ECOG, se considera como un factor pronóstico al valorar la repercusión que la enfermedad produce en el estado general del paciente. Todas las situaciones clínicas previas que puedan influir en la morbimortalidad y tolerancia al tratamiento se consideran como factores pronósticos (enfermedades cardiovasculares, diabetes, hepatitis, etc.). Los linfomas que aparecen en situaciones de inmunodeficiencia tienen un curso más agresivo y peor pronóstico; prueba de ello son los LNH asociados al síndrome de inmunodeficiencia adquirida (SIDA).

- Dentro de los factores pronósticos dependientes del tumor se considera el subtipo histológico, el inmunofenotipo, las alteraciones citogenéticas, actividad proliferativa y la extensión de la enfermedad. El subtipo histológico, el patrón de infiltración y aspectos citológicos de las células así como su diferenciación se consideran como factores pronósticos; aunque el pronóstico de las diferentes entidades no son sustancialmente diferentes. El inmunofenotipo demuestra un peor pronóstico del linaje T frente al B. Las anomalías citogenéticas están presentes en la mayoría de los LNH, la presencia de alteraciones cromosómicas y el número de éstas reviste peor pronóstico.

  La extensión de la enfermedad, definida como la cantidad del tumor al diagnóstico, reviste de importancia prónostica en los LNH. Los siguientes parámetros son considerados como factores pronósticos: el estadio clínico, número y localización de áreas ganglionares y extraganglionares afectas, tamaño del tumor (tamaño del diámetro mayor superior a $10cm$, "masa Bulky"), carga tumoral (número de regiones ganglionares extensas y el número de localizaciones extraganglionares).

Otras variables con significado pronóstico son: presencia de síntomas B (fiebre, sudoración nocturna y pérdida de peso), hemoglobina baja, leucocitos elevados, deshidrogenasa láctica y $\beta$2-microglobulina elevada.

# Capítulo 3
# Conceptos básicos en análisis de supervivencia

En este capítulo se presentan algunos conceptos básicos en análisis de supervivencia, que serán utilizadas en los siguientes capítulos como son: datos de supervivencia, funciones de distribución, procesos de conteo y suavizamiento con splines.

## 3.1. Datos en análisis de supervivencia

En muchos estudios clínicos los investigadores pueden estar interesados en analizar el tiempo de supervivencia y las características clínicas de los pacientes que pueden estar relacionados con el tiempo de supervivencia. En este contexto, hay tres aspectos característicos en el análisis de los datos de supervivencia, que son:

- El tiempo de seguimiento hasta la ocurrencia del evento, llamado tiempo de supervivencia.
- La censura o censuramiento, que se origina debido a estudios que son terminados antes de los resultados de todas las unidades conocidas, y
- La presencia de las variables explicativas, llamadas covariables.

### 3.1.1. Tiempo de supervivencia.

Se denomina tiempo de supervivencia al tiempo transcurrido entre la fecha de inicio (ingreso al estudio, inicio de tratamiento, etc.) y la fecha de ocurrencia de un evento de interés (recaída, progresión, muerte, etc.) o la fecha en que finaliza el estudio. En general, el tiempo de supervivencia es un proceso continuo, la longitud de la duración puede medirse utilizando un número real no negativo.

Como un término genérico, el tiempo desde la iniciación de un evento (nacimiento, diagnóstico, inicio de tratamiento, etc) hasta la ocurrencia de un evento de interés (recaída, progresión, muerte, etc.) se denomina como tiempo de supervivencia, aún cuando el evento final es algo diferente de la muerte. Algunos ejemplos:

- Tiempo desde el nacimiento hasta la muerte.
- Tiempo desde el nacimiento hasta el diagnóstico de cáncer.
- Tiempo transcurrido desde la aparición de la enfermedad hasta la muerte.
- Tiempo transcurrido desde la respuesta clínica a la recaída.

En general, para fines de nuestra aplicación, se considera el análisis de supervivencia clásico que se centra en el tiempo hasta la ocurrencia de un simple evento (muerte del paciente) para cada individuo, o más exactamente el tiempo transcurrido desde inicio hasta la ocurrencia del evento muerte.

### 3.1.2. Censura o Censuramiento.

Normalmente, los estudios de supervivencia tienen una duración predeterminada, por lo que no todos los sujetos en seguimiento habrán fallado a su finalización. Por lo tanto, el investigador sabrá que un cierto número de individuos han "sobrevivido"durante el periodo de tiempo, pero desconocerá el momento exacto en que hubiera fallado si el estudio si hubiera prolongado de forma indefinida. A este tipo de datos se llaman observaciones censuradas.

Se dice que las observaciones están censuradas cuando contienen información parcial sobre los tiempos de supervivenvia durante un periodo de seguimiento. La información parcial, ocurre debido a causas como:

- Retiro del estudio por causas ajenos al evento de interés
- Pérdida de acompañamiento, o
- Finalización del estudio.

En general, el término de censura hace referencia a un tipo de pérdida de información en situaciones en las que la variable de interés es el tiempo de supervivencia. La censura surge en las ocasiones en las que hay individuos de la muestra para los que no se conoce exactamente su tiempo de supervivencia, sino que únicamente se sabe que éste ha ocurrido dentro de un cierto intervalo de tiempo. De esta forma se puede considerar tres tipos de censura: censura por la derecha, por la izquierda y censura en un intervalo.

Se dice censura por la derecha, cuando en el momento en que finaliza el estudio hay sujetos para los que no se conoce el instante exacto de falla, sino que solamente se sabe que es posterior a un momento dado. Censura por la izquierda, cuando el momento exacto en que

ocurrió la falla es desconocido, tan sólo se sabe que ha ocurrido antes de que el sujeto se incluya en el estudio. Y censura en un intervalo cuando el evento de interés no se puede observar exactamente y sólo se sabe que ha ocurrido en un cierto intervalo de tiempo.

En los casos de censura por la derecha se tiene tres tipos de censura: censura de tipo I (censuramiento por tiempo), tipo II (censuramiento por fallas) y tipo III (censuramiento aleatorio). Si un estudio termina en un tiempo pre-establecido y algunos de los tiempos de supervivenvia son no observados, tenemos censura tipo I. En el caso que el estudio termina después de la ocurrencia de una determinada cantidad pre-establacida de eventos, tenemos censura de tipo II.

La censura tipo III o aleatoria surge de manera natural en las investigaciones biomédicas debido a que los pacientes entran al estudio en tiempos diferentes y de manera aleatoria, y que cada paciente tiene un modo propio de censura, debido a cualquiera de las causas descritas (retiro, pérdida de seguimiento, finalización del estudio), de modo que los tiempos de censuramiento son también aleatorias.

Para propósitos del presente trabajo nos enfocaremos en datos con censura por la derecha y de tipo aleatorio.

### 3.1.3. Covariables.

Además de los datos de supervivencia (tiempo de supervivencia y la variable indicadora de censura), también se pueden observar otros datos, variables que representan la heterogeneidad existente en la población, tales como, la edad, el género, la hemoglobina, estadio clínico, etc. Estas variables son conocidas como variables explicativas o covariables y son muy frecuente en muchos estudios clínicos.

En general, las covariables son variables independientes y observables. Según la escala de medición, se pueden clasificar en variables cuantivativas y cualitativas, y segun la evolución en el tiempo en variables fijas o tiempo-dependientes.

#### *3.1.3.1. Clasificacíon según la escala de medición.*

Las variables cuantitativas son variables que se pueden medir expresándose numéricamente. Estas variables pueden ser de dos tipos: Continuas, cuando admiten tomar cualquier valor dentro de un rango numérico (ejemplo: edad, peso, talla, tamaño del tumor, hemoglobina, etc). Discretas, si solamente toman valores enteros, por lo que no admiten los valores intermedios en un rango dado (ejemplo: número de hijos, número de ganglios, etc).

Las variables cualitativas son variables que representan distintas cualidades de un individuo; estas cualidades se denominan atributos o categorías, y la medición de éstas variables

consiste en la clasificación de dichos atributos. En el proceso de medición de las variables cualitativas, se pueden utilizar dos escalas, la ordinal y la nominal. En la escala ordinal, la clasificación de las categorías presentan un orden natural (ejemplo: grupos de edad, estado de performance, estadio clínico, etc). En la escala nominal, las categorías de la variables no se clasifican de acuerdo a un criterio de orden tanto inherente como jerárquico (ejemplo: género, estado civil, etc). Dependiendo de los valores que tome una variable cualitativa, éstas pueden ser dicotómicas o bien politómicas.

*3.1.3.2. Clasificacíon según la evolución en el tiempo.*

Las variables fijas, son variables cuyos valores no varían durante la evolucion del estudio; es decir el valor de estas variables no cambian durante el periodo de seguimiento. El valor o el atributo de estas variables al inicio del estudio es la misma en cualquier momento del tiempo (ejemplo: género, raza, tipo de Rh, etc).

Una complicación que puede ocurrir en el análisis de supervivencia es observar variables que pueden variar en el tiempo, denominadas variables tiempo-dependientes. Los valores de estas variables no es lo mismo al final que al inicio del estudio. (ejemplo: edad, estado de la enfermedad, respuesta al tratamiento, modificación de la dosis de un determinado medicamento a lo largo del tratamiento, etc.)

## 3.2. Funciones del tiempo de supervivencia

En análisis de supervivencia existen dos funciones de gran interés: la función de supervivencia y la función de riesgo; las cuales, son descritas brevemente en esta sección. Para comenzar con el modelado de datos de supervivencia se partirá de la suposición de que la población es homogénea.

Sea $T$ el tiempo de supervivencia, una variable aleatoria positiva ($T > 0$) con función de densidad $f(t)$ y función de distribución $F(t)$. La función de densidad es definida como el límite de la probabilidad de un individuo de fallecer en un intervalo de tiempo $[t, t + \triangle t)$ por unidad de tiempo, y es expresada por

$$f(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T < t + \triangle t)}{\triangle t} \tag{1}$$

y la función de distribución es definida como la probabilidad de fallecer de un invididuo antes de un tiempo $t$, y es expresada por

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

Además de $f$ y $F$, en el análisis de supervivencia, la distribución del tiempo de supervivencia puede ser caracterizada por otras funciones equivalentes: la función de supervivencia y la función de riesgo.

La función de supervivencia es definida como la probabilidad de que un individuo sobreviva por lo menos un determinado tiempo $t$. Esta función es decreciente con un valor 1 para $T = 0$ y cero para $T = \infty$, y es expresada como

$$S(t) = P(T > t) = 1 - F(t) = 1 - \int_t^\infty f(s)ds. \tag{2}$$

La función de riesgo se define como la probabilidad condicional de que un sujeto muera en un intervalo de tiempo $(t, t + \Delta t)$ dado que ya sobrevivió por lo menos un tiempo $t$, interpretado como la tasa instantánea de falla, y es expresada como

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t / T > t)}{\Delta t} \tag{3}$$

De la expresión (1) y (2) la función de riesgo es, $\lambda(t) = f(t)/S(t)$ y la función de riesgo acumulado se denota como $H(t) = \int_t^\infty \lambda(s)ds$; por lo tanto, la función de supervivencia puede ser calculada a partir de la función de riesgo por

$$S(t) = \exp\left(-\int_t^\infty \lambda(s)ds\right) = \exp(-H(t))$$

En consecuencia, en el análisis de los datos de supervivencia, la descripción y modelado de los tiempos de supervivencia se puede realizar con cualquiera de estas funciones, denominadas funciones equivalentes.

## 3.3. Función de máxima verosimilitud

Una de las partes fundamentales de todo procedimiento estadístico, es la estimación de los parámetros del modelo estadístico basado en una muestra. El procedimiento de estimación en una muestra no censurada no es complicado; sin embargo, en una muestra censurada el procedimiento de estimación de los parámetros esta sujeto a un factor o indicador de censura o censuramiento.

En las investigaciones biomédicas, el censuramiento surge de manera natural, debido a que los pacientes entran al estudio en tiempos diferentes, de manera que cada paciente tiene un modo propio de censuramiento por la derecha, debido a cualquiera de las tres causas, de modo que los tiempos de censuramiento son también aleatorios.

En esta sección se describe brevemente la función de verosimilitud para el procedimiento de estimación de los parámetros del modelo o la función de supervivencia bajo censuramiento por la derecha y de manera aleatoria, mediante el método de máxima verosimilitud.

Sea $Y$ el tiempo de supervivencia y $C$ el tiempo de censuramiento asociado, con función de densidad y función de supervivencia, ($f_T(t)$, $S_T(t)$) y ($g_C(t)$ y $1-G_C(t)$), respectivamente. Bajo un mecanismo de censura no-informativo, se supone que $Y$ y $C$ son independientes. También se asume que $G(t)$ no depende de ninguno de los parámetros de $S(t)$, por lo que no aporta información alguna para la distribución del tiempo de supervivencia. En este modelo de censura, lo que se observa por unidad muestral es el par aleatorio $(T, \delta)$ definido como

$$T = \min(Y, C),$$

y

$$\delta = I_{[Y \leq C]} = \begin{cases} 1, & \text{si } T \text{ es no censurado} \\ 0, & \text{si } T \text{ es censurado} \end{cases}$$

donde $\delta$ es la variable indicadora de censura.

Sean los tiempos de supervivencia observados en una muestra de $n$ individuos que consiste de los pares $(t_1, \delta_1), \ldots, (t_n, \delta_n)$. La función de verosimilitud es dada por

$$L = \prod_{i=1}^{n} [f(t_i)(1 - G_i)]^{\delta_i} [g_i S(t_i)]^{1-\delta_i}. \tag{4}$$

Debido a que el tiempo de censuramiento es no informativo, la función de verosimilitud se reduce en términos de $f(t)$ y $S(t)$ a:

$$L = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}. \tag{5}$$

En consecuencia, reemplazando las funciones respectivas en la expresión (5), se pueden obtener los estimadores de los parámetros del modelo de distribución para variable aleatoria $T$, la función de supervivencia o las funciones equivalentes.

## 3.4. Procesos de conteo

En el análisis de supervivencia, el enfoque del análisis está en la observación de la ocurrencia de eventos sobre el tiempo. Dichas ocurrencias constituyen procesos puntuales. Estos procesos pueden ser descritos como el conteo del número de eventos que se van presentando durante el tiempo, lo que lleva al término de " procesos de conteo". Algunos ejemplo pueden ser:

- Contar el número de veces que una persona se despierta durante la noche.
- Contar las muertes en un grupo de pacientes con tratamiento en un ensayo clínico.

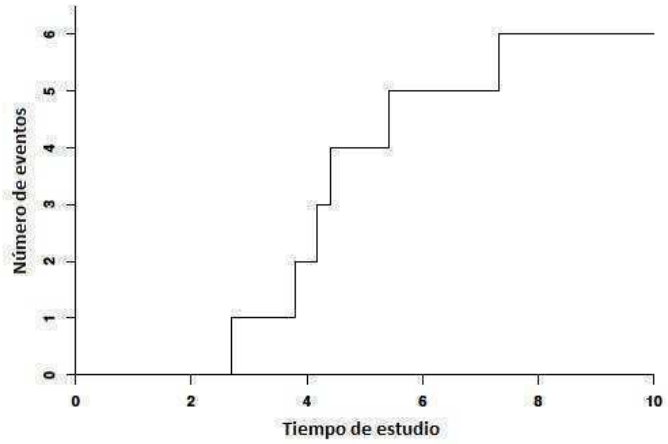
En esta sección se describe brevemente los términos básicos relacionados al proceso de conteo, como son proceso de conteo y la martingala, conceptos que son utilizados para realizar el diangóstico tiempo-dependiente del modelo y la forma funcional del efecto de las covariables basado en los residuos. En el modelo de Cox, el diagnóstico del modelo se basa en los residusos martingala para determinar la forma funcional de los residuos martingala en relación a las covariables que sugieriese falta de ajuste.

### 3.4.1. Conceptos básicos relacionados al procesos de conteo.

Los tiempos de supervivencia pueden ser representados a través de ciertos procesos estocásticos. Los datos en sí pueden ser descritos como un proceso de conteo, el cual es simplemente una función aleatoria del tiempo $t$, $N(t)$. Esta función es cero en el tiempo inicial y constante en el tiempo, excepto en el tiempo donde ocurre el evento, donde hace un salto de tamaño 1.

Considerando un solo evento para un tiempo $t$ dado. Sea $N(t)$ el número de eventos que han ocurrido hasta el tiempo $t$, entonces $N(t)$ es un proceso de conteo.

Sean los datos de 10 pacientes de un estudio clínico hipotético, escrito en orden creciente: 2.70, 3.50+, 3.80, 4.19, 4.42, 5.43, 6.32+, 6.46+, 7.32,8.11+. El proceso de conteo correspondiente para estos datos se ilustra en la Gráfica 1.




Gráfica 1. Ilustración de un proceso de conteo.

En la Gráfica 1 se observa que el proceso da saltos de una unidad en cada evento observado en el tiempo, es constante entre los eventos y es continua por la derecha.

Así, los tiempos de supervivencia pueden ser representados a través de ciertos procesos estocásticos. En los párrafos siguientes se describen los términos básicos de uso común en la metodología de los procesos de conteo.

Sea $t$ una variable tiempo, tal que, $N(t)$ es definida como el número de eventos que han ocurrido hasta el tiempo $t$, entonces $N(t)$ es un proceso de conteo. Es decir, para los puntos $t_j$ aleatoriamente dispuestos a lo largo de una línea, el proceso de conteo $N(t)$ da el número de puntos observados en el intervalo $(0, t]$:

$$N(t) = \#\{t_j\,, 0 < t_j \leq t\}$$

donde # representa la cardinalidad (número de elementos) de un conjunto.

La historia, $H_t$, consiste en determinar las variables hasta el tiempo $t$ que son necesarios para describir la evolución del proceso de conteo. La historia se llama a menudo la filtración $(\mathfrak{F}_t)$ en la literatura de procesos de conteo (Ver Andersen et al. (1993), Touboul y Faugeras (2007), para una definición rigurosa de los conceptos).

Para el proceso $N$ y historia $H_t$ al tiempo $t$, la función de intensidad se define como:

$$\lambda(t|H_t) = \lim_{h\to 0} \frac{P\{\text{evento} \in (t, t+h]|H_t\}}{h};$$

para un $h$ pequeño se tiene:

$$P\{evento \in (t, t+h]|H_t\} \approx \lambda(t|H_t)h$$

El proceso de intensidad acumulada se define como,

$$\Lambda(t) = \int_0^t \lambda(s)ds.$$

Y el proceso de conteo martingala, llamada "martingala"se define como,

$$M(t) = N(t) - \Lambda(t)$$

En general, hay una teoría matemática rigurosa para los procesos de conteo, los cuales son extremadamente útiles para el análisis estadístico de los datos de supervivencia y datos de eventos históricos. La razón de usar procesos de conteo y martingalas es porque éstos métodos proporcionan formas directas de estudiar las propiedades de muestras grandes de los estimadores.

### 3.4.2. Procesos de conteo y datos censurados.

Una aproximación alternativa para desarrollar los procedimientos de inferencia para datos censurados es utilizar la metodología de procesos de conteo. Esta aproximación fue desarrollada por Aalen (1975), quien combina elementos de integración estocástica, teoría de martingalas y teoría de procesos de conteo dentro de una metodología que permite desarrollar fácilmente las técnicas de inferencia para el análisis de supervivencia con datos censurados (Andersen et al. (1993), Fleming y Harrington (1991)).

Un proceso de conteo $\{N(t), t \geq 0\}$ es un proceso estocástico con las siguientes propiedades: $N(0) = 0$, $N(t) \geq 0$, $N(t) < \infty$, $N(t)$ es un número entero con probabilidad 1 y las trayectorias de la muestra de $N(t)$ son continuos por la derecha y constantes entre eventos y saltos de tamaño 1.

Dada una muestra aleatoria con censura por la derecha, los procesos

$$N_i(t) = I\{T_i \leq t, \delta_i = 1\},$$

son procesos de conteo, los cuales son cero hasta que el individuo $i$ presenta el evento y entonces tiene un salto de tamaño 1.

El proceso

$$N(t) = \sum_{i=1}^{n} N_i(t) = \sum_{t_i \leq t} \delta_i,$$

también es un proceso de conteo. Este proceso simplemente cuenta el número de muertes en la muestra al tiempo anterior o igual a $t$.

El proceso de conteo da información acerca de cuándo ocurren los eventos. Además de conocer esta información, podemos tener información adicional sobre los sujetos en estudio. Para los datos censurados por la derecha, la información incluye el conocimiento de quién ha sido censurado o quién murió antes del tiempo $t$. En algunos casos la información puede incluir datos de covariables fijos en el tiempo (ejemplo, educación,sexo, tratamiento, etc.) y/o covariables que pueden variar en el tiempo (ejemplo, edad, carga tumoral, etc.), todos antes del tiempo $t$ (denominada historia $(H_t)$ o filtración del proceso de conteo $(\mathfrak{F}_t)$)).

En el caso de datos con censura por la derecha, la filtración al tiempo $t$, $\mathfrak{F}_t$, consiste en el conocimiento de los pares $(T_i, \delta_i)$ siempre que $T_i \leq t$ para los individuos que continúan en el estudio al tiempo $t$. Se denota la filtración a un instante antes de $t$ por $\mathfrak{F}_{t-}$. La filtración $\{\mathfrak{F}_t, t \geq 0\}$ para un problema dado depende de las observaciones del proceso de conteo.

Para datos censurados por la derecha, si los tiempos de muerte $T_i$ y los tiempos de censura $C_i$ son independientes, entonces el cambio de un evento al tiempo $t$, dada la historia antes del tiempo $t$, está dado por

$$\begin{aligned} &P[t \leq T_i \leq t + dt, \delta_i = 1 | \mathfrak{F}_{t-}] \\ &= \begin{cases} P(t \leq T_i \leq t + dt, C_i > t + dt_i | T_i \geq t, C_i \geq t) = h(t)dt & \text{si} \quad T_i \geq t \\ 0 & \text{si} \quad T_i < t \end{cases} \end{aligned} \tag{6}$$

Para un proceso conteo dado, $dN(t)$ se define como el cambio en el proceso $N(t)$ sobre un intervalo de tiempo corto $[t, t + dt)$. Esto es,

$$dN(t) = N[(t + dt)^-] - N(t^-), \tag{7}$$

donde $t^-$ es el tiempo justo antes de $t$).

En datos censurados por la derecha (asumiendo no empates), $dN(t)$ es uno si la muerte ocurre en $t$ o 0 en otro caso.

Si definimos el proceso $Y(t)$ como el número de individuos con un tiempo de estudio $T_i \geq t$, entonces, usando (6),

$$\begin{aligned} E(dN(t)|\mathfrak{F}_{t^-}) =& E[\text{número de observaciones con} \\ & t \leq T_i \leq t + dt, C_i > t + dt_i | \mathfrak{F}_{t^-}] \\ =& Y(t)h(t)dt \end{aligned}$$

El proceso $\lambda(t) = Y(t)h(t)$ es llamado proceso de intensidad (intensity process) de un proceso de conteo. $\lambda(t)$ es en sí un proceso estocástico que depende de la información contenida en la historia, $\mathfrak{F}_t$ a través de $Y(t)$.

El proceso estocástico $Y(t)$ es el proceso que proporciona el número de individuos en riesgo en un momento dado del tiempo $t$, y junto con $N(t)$, es una cantidad fundamental en los métodos presentados.

Para el caso continuo, el proceso de intensidad acumulado se define como $\Lambda(t) = \int_0^t \lambda(s)ds$, $t \geq 0$; el cual, tiene la siguiente propiedad: $E(N(t)|\mathfrak{F}_{t^-}) = E(\Lambda(t)|\mathfrak{F}_{t^-}) = \Lambda(t)$. Esta última igualdad se cumple porque una vez que conocemos la historia justo antes de $t$, el valor de $Y(t)$ es fijo y entonces $\Lambda(t)$ es no aleatoria.

El proceso estocástico $M(t) = N(t) - \Lambda(t)$ es llamado proceso de conteo martingala. Este proceso tiene la propiedad que el incremento de este proceso tiene un valor esperado, dado el pasado estricto, $\mathfrak{F}_{t^-}$, iguales a cero, esto es,

$$E(dM(t)|\mathfrak{F}_{t^-}) = E[dN(t) - d\Lambda(t)|\mathfrak{F}_{t^-}] = 0 \tag{8}$$

La expresión (8) es muy interesante porque es precisamente la definición intuitiva de una martingala.

En general, sean $n$ sujetos independientes que son observados durante un período de tiempo $[0, t)$. Para cada sujeto un proceso de conteo, $N_i(t)$, que da el número de eventos ocurridos antes del tiempo $t$ es observado junto con posible información adicional en términos de las covariables $X_i$ $p$-dimensional.

Modelos para datos de supervivencia, o más generalmente datos de procesos de conteo, son muy convenientemente formulados a través de la *función de intensidad del proceso* $\lambda(t)$, que se define como (Scheike, 2004)

$$\lambda_i(t) = \lim_{h \to 0} \frac{P(N_i(t+h) - N_i(t) \geq 1|\mathfrak{F}_{t^-})}{h},$$

La *función de intensidad acumulada* se define como

$$\Lambda_i(t) = \int_0^t \lambda_i(s)ds.$$

Así mismo, dada $N_i(t)$ y $\Lambda_i(t)$, el proceso martingala se expresa como

$$M_i(t) = N_i(t) - \Lambda_i(t)$$

En adelante se considera $N(t) = (N_1(t), \ldots, N_n(t))$ como un proceso de conteo $n$-dimensional, $\Lambda(t) = (\Lambda_1(t),\ldots,\Lambda_n(t))$ su compensador, $\lambda(t) = (\lambda_1(t), \ldots, \lambda_n(t))$ proceso de intensidad $n$-dimensional, y $M(t) = (M_1(t),\ldots, M_n(t))$ la martingala $n$-dimensional.

## 3.5. Suavizamiento con splines

Los modelos paramétricos constituyen un método eficiente cuando se tiene información del modelo subyacente a las variables y sólo resta por determinar un número finito de parámetros; sin embargo, una fuente de error puede ser elegir una familia paramétrica no adecuada. En estos casos podemos utilizar los métodos no-paramétricos, que además de permitir graduar las probabilidades brutas que no siguen una forma paramétrica clara, pueden utilizarse para proporcionar una prueba de diagnóstico de los modelos paramétricos o simplemente para explorar los datos. Las funciones que habitualmente se estiman son: la función de densidad, la función de regresión y sus derivadas.

La teoría de los métodos no-paramétricos desarrolla procedimientos de inferencia estadística, que no realizan una suposición explícita con respecto a la forma funcional de la distribución de probabilidad de las observaciones de la muestra. Si bien en la estadística no-paramétrica también aparecen modelos y parámetros, ellos están definidos de una manera más general que en su contraparte paramétrica. En este contexto, las técnicas de suavizado tienen en la actualidad un papel muy relevante. Esta popularidad se debe en parte a la complejidad de los datos que se generan y que hacen que un modelo paramétrico sea inviable. Además, los avances informáticos han reducido el coste computacional que supone utilizar modelos no-paramétricos.

Supongamos que tenemos observaciones $(y, x)$ de un modelo estadístico del tipo $y = f(x)+e$, donde $e$ es una variable aleatoria con media cero y varianza constante. Si $f(x) = a + bx$, el modelo se reduce a una regresión lineal simple. En situaciones de una relación no-lineal, podría utilizarse un ajuste polinomial, aunque el ajuste polinomial es sensible a outliers; sin embargo, los outliers podrían tener un efecto más local cuando se utiliza aproximación mediante polinomio por trozos.

En general, sea $f(\cdot)$ una función desconocida, denotado como $s(x)$ de aquí en adelante, es decir la función no asume una forma funcional paramétrica. Una solución estadística a este problema se puede realizar utilizando modelos de regresión no-paramétrica. Desde el enfoque no-paramétrico, la estimación de la función $s(\cdot)$ se podría realizar mediante distintos métodos divididos en dos grandes grupos: regresión tipo kernel y regresión con splines.

Dentro de las técnicas de suavizado basadas en los splines hay dos familias: a) suavizamiento splines (smoothing splines) y b) regresión splines (regression splines). El suaviamiento splines utiliza tantos parámetros como observaciones, lo que hace que su implementación no sea eficiente cuando el número de datos es muy elevado. La regresión splines puede ser ajustada mediante mínimos cuadrados una vez que se han seleccionado el número de nodos, donde

la selección de los nodos se hace mediante algoritmos bastante complicados. En cambio los splines con penalización (penalized splines) combinan lo mejor de ambos enfoques.

En esta sección se describe brevemente los conceptos básicos relacionados con la función splines, bases B-splines y splines penalizados (P-splines).

### 3.5.1. Función splines.

Los splines fueron implementados en estadística por Wahba (1990), pero sus orígenes se remontan a 1923 gracias a la teoría desarrollada por Whittaker. Un spline es simplemente una curva. En matemática, un spline es una función especial definida parcialmente por polinomios.

En general, un spline es un polinomio por trozos con partes definidas por una secuencia de nodos $\xi_1 < \xi_2 < \ldots < \xi_K$ con la condición de continuidad en la función y sus derivadas en los puntos donde se unen los trozos, de tal manera que las partes se unen sin problemas en los nodos.

La función $s : [a, b] \to R$ es una función spline (o un spline) de grado $d$ con nodos $\xi_1, \ldots, \xi_k$, si se verifica las siguientes condiciones:

- $a < \xi_1 < \ldots < \xi_K < b$ ($\xi_0 = a$, $\xi_{K+1} = b$)
- En cada intervalo $[\xi_j, \xi_{j+1}]$ ($j = 0, \ldots, k$), $s(x)$ es un polinomio de grado $d$
- La función $s(x)$ tiene $(d-1)$ derivadas continuas en $[a, b]$

Sea $s : [p; a = \xi_0, \xi_1, \ldots, \xi_k, \xi_{k+1} = b]$ el conjunto de los splines de grado $d$ con nodos $\xi_1, \ldots, \xi_k$, definido en $[a, b]$. $s : [p; a = \xi_0, \xi_1, \ldots, \xi_k, \xi_{k+1} = b]$ es un espacio vectorial de dimensión $d + k + 1$.

Para un spline de grado $d$ se requiere usualmente de los polinomios y sus primeras $d - 1$ derivadas de acuerdo a los nodos, por lo que las $d - 1$ derivadas son continuas.

Por ejemplo, un spline de grado $d$ se puede representar como una serie de potencias:

$$s(x) = \sum_{j=0}^{d} \beta_j x^j + \sum_{j=1}^{k} \gamma_j (x - \xi_j)_+^d \tag{9}$$

donde la notación

$$(x - \xi_j)_+ = \begin{cases} x - \xi_j, & x > \xi_j \\ 0, & \text{en otro caso} \end{cases}$$

De la expresión (9), un spline lineal con 1 nodo se expresa como:

$$s(x) = \beta_0 + \beta_1 x + \gamma(x - \xi)_+,$$

y un spline cúbico con $k$ nodos se expresa como:

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{k} \gamma_j (x - \xi_j)_+^3.$$

En general, una funciones splines es una combinación lineal o expansión lineal de las funciones bases. Para una spline de grado $d$ y $k$ la función se expresa como

$$s(x) = \sum_{j=0}^{d+k} \theta_j h_j(x, \xi), \tag{10}$$

donde cada $h_j$ es una función del predictor $x$, y el término lineal aquí se refiere a la acción del parámetro $\theta$.

La expresión (10) incluye a una familia lineal y expanciones polinomiales, sobre todo a una gran variedad de modelos flexibles. Una vez que se determinan las funciones bases, los modelos son lineales en estas nuevas variables y el ajuste se procede como en una regresión lineal.

### 3.5.2. Bases y nodos.

Hay muchas maneras de calcular la base para la regresión, de hecho las más conocidas son:

- Bases de potencias truncadas (Ruppert et. al (2003))
- Bases B-splines (De Boore (1977) y Dierckx (1993))
- Bases "thin plate regression splines"(Green y Silverman (1994), Wood (2003))

De las tres bases, las bases B-splines son la más recomendadas ya que son numéricamente más estables que otras bases (como es el caso de los polinomios truncados). De manera que, en la sección siguiente se describen las bases B-splines con más detalle, debido a que esta base es utilizada en la aproximación del efecto de las covariables en el modelo de Cox con regresión splines y modelo de Cox con P-splines.

### 3.5.3. Bases de potencia truncada.

Una función de potencia truncada de grado $d$ para un nodo $\xi_1$ es una función definida por

$$\psi_1(x) = \begin{cases} (x-\xi_1)^d, & x \geq \xi_1 \\ 0, & \text{en otro caso} \end{cases}$$

Como una función de $x$, esta función toma el valor 0 a la izquierda de $\xi_1$, y toma el valor $(x-\xi_1)$ a la derecha de $\xi_1$.

El nombre se deriva del hecho de que estas funciones son funciones de potencias desplazadas que se truncan a cero a la izquierda del nodo. Estas funciones son funciones polinomiales a trozos con dos partes cuya función y sus derivadas hasta $d-1$ son iguales a cero en el nodo. Por tanto, estas funciones son splines de grado $d$. Es fácil ver que estas funciones son linealmente independientes. Sin embargo, no forman una base, porque una base requiere $d+k+1$ funciones.

La forma usual para añadir $d+1$ funciones bases adicionales es el uso de los polinomios $1, x, x^2, \ldots, x^d$. Estos $d+1$ funciones junto con las $k$ funciones de potencia truncada $\psi_j(x)$, $i = 1, 2, \ldots, k$ forman las bases de potencias truncadas.

Es decir, para $k$ nodos en el intervalo $[a, b]$, los elementos de las bases de potencias truncados de grado $d$ viene dada por:

$$1, x, x^2, \ldots, x^d, (x-\xi_1)_+^d, \ldots, (x-\xi_k)_+^d, \tag{11}$$

Las función ${(x-\xi)_+}^d$ tienen $d-1$ derivadas continuas, de modo que cuanto mayor sea $d$ más suave son las funciones en la base.

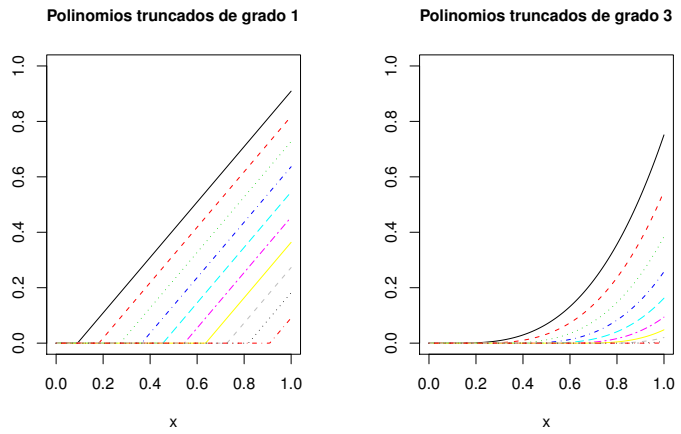
De la expresión (11) una función spline cúbico con bases de pontencia truncada y $k$ nodos es expresado como

$$s(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \sum_{j=1}^{k} \alpha_j (x-\xi_j)_+^3 \tag{12}$$

La Gráfica 2 muestra tales funciones para $d = 1$ y $d = 3$ con 10 nodos equiespaciados.

La principal ventaja de la función base de potencia truncada es la simplicidad de su construcción y la facilidad de interpretación de los parámetros de un modelo que corresponde para estas funciones base. Sin embargo, hay dos puntos débiles cuando se utiliza esta base para la regresión.

Estas funciones crecen rápidamente sin límite a medida que $x$ incrementa, resultando en problemas de precisión numérica, cuando los datos $x$ abarcan un amplio rango de valores. Además, muchos o incluso la totalidad de estas funciones base pueden ser distintos de cero




GRÁFICA 2. Bases de polinimios truncados de grado 1 (izquierda) y grado 3 (derecha) definidos en [0,1] con 10 nodos equi-espaciados, respectivamente.

cuando se evalúa en un valor $x$. Esta debilidad puede ser abordado mediante el uso de bases B-spline.

### 3.5.4. Bases B-splines.

B-splines fue introducido por Schoenberg (1946). Muchas de sus propiedades algebraicas pueden encontrarse en Curry y Schoenberg (1966). Las referencias básicas son De Boor (1977) y Dierckx (1993). Un B-spline está formado por trozos de polinomios conectados entre si.

Un ejemplo de las bases B-splines se muestra en la Gráfica 3. En la parte izquierda aparece bases B-splines de grado 1 que están formados por dos trozos de polinomio lineal que se unen en un nodo, donde cada uno está basado en tres nodos. En la parte derecha aparece B-splines de grado tres, formados por 4 trozos de polinomios unidos entre si (tres nodos) basado en cinco nodos. Todas las funciones de la base tienen la misma forma, pero están desplazadas horizontalmente (el desplazamiento es una función de la distancia entre los nodos).

Los B-splines son una base del espacio vectorial de funciones splines de grado $p$ con nodos $t_1, \ldots, t_k$ definidos en $[a, b]$, $S : [p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$, que representan ventajas computacionales respecto a las bases de potencias truncadas.

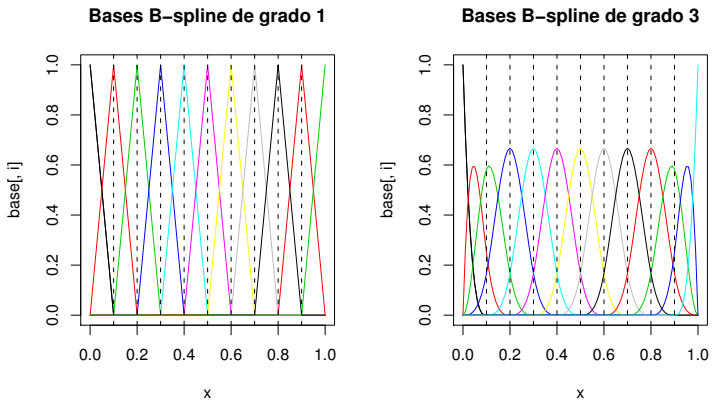
Las bases B-splines se definen recursivamente. Además de los $k$ nodos se definen $2M$ nodos auxiliares: $\tau_1 \leq \ldots \leq \tau_M \leq t_0, t_{k+1} \leq \tau_{k+M+1} \leq \ldots \leq \tau_{k+2M}$. La elección de los nodos es arbitraria y se puede hacer $\tau_1 = \ldots = \tau_M = t_0, t_{k+1} = \tau_{k+M+1} = \ldots = \tau_{k+2M}$. Se renombra los nodos originales como $\tau_{M+j} = t_j$, $j = 1, \ldots, k$.

La base B-spline de orden $m = 1$ es: $B_{j,1} = I_{[\tau_j, \tau_{j+1}]}$, $j = 1, \ldots, k + 2M - 1$

Para $m = 2, \ldots, M$, los B-splines de orden $m$ se definen como

$$B_{j,m} = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_{j,m-1} + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1,m-1}. \tag{13}$$

En la Gráfica 3 se muestra las 13 funciones que forman las bases B-splines de grado 1 (izquierda) y 3 (derecha) definidos en [0; 1] con nueve nodos equi-espaciados en 0,1, 0,2, . . . , 0,9, respectivamente. Como se puede observar las bases B-spline de grado 1 están formados por dos trozos de polinomios lineales que se unen en un nodo interno y las bases B-spline de grado 3 están formados por cuatro trozos de polinomios unidos en tres nodos internos.




GRÁFICA 3. Bases B-splines de grado 1 (izquierda) y 3 (derecha) definidos en [0,1] con 9 nodos equi-espaciados, respectivamente.

Tres propiedades importantes de los B-splines (deBoor 1978) son:

- $B_j(x) = 0$, si $x$ no pertenece $[t_j, t_{j+m}]$
- $B_j(x) \geq 0$, si $x$ pertenece $[t_j, t_{j+m}]$
- $\sum B_j(x) = 1$

De la expresión (13) una función splines cúbico con bases B-spline y $k$ nodos es expresado como

$$s(x) = \sum_{j=1}^{3+k+1} \theta_j B_j(x, \xi). \tag{14}$$

### 3.5.5. P-splines.

Eilers y Marx (1996) introducen el término $P$-splines, llamado splines penalizado (Ruppert y Carroll, 2000) o pseudo-splines (Hastie, 1996) y son una extensión de los B-splines y comparten muchas de las propiedades.

Los P-splines es un término intermedio entre el suavizamiento splines y regresión splines, de hecho combinan lo mejor de ambos enfoques. Los P-splines utilizan menos parámetros que los splines de suavizamiento, pero la selección de los nodos no es tan determinante como en los splines de regresión. Son splines de rango bajo, el número de nodos es mucho menor que la dimensión de los datos, al contrario de lo que ocurre en el caso de los splines de suavizamiento. El número de nodos, en el caso de los P-splines, no supera los 40, lo que hace que sean computacionalmente más eficientes, sobre todo cuando se trabaja con gran cantidad de datos. Además, la introducción de penalizaciones relaja la importancia de la elección del número y la localización de los nodos, cuestión que es de gran importancia en los splines de rango bajo sin penalizaciones (Rice y Wu, 2001).

La novedad es que los autores proponen el uso de B-splines simétrico y penalizan estos, no en la segunda derivada, sino en las diferencias entre los coeficientes splines adyacentes. Este tipo de penalización es más flexible ya que es independiente del grado del polinomio utilizado para construir los B-splines, un criterio que es fácil de implementar, que resulta estrechamente relacionado con la usual penalización.

Sean los términos flexibles expresados como:

$$f(x) = \sum_{m=1}^{M} \alpha_m B_m(x), \tag{15}$$

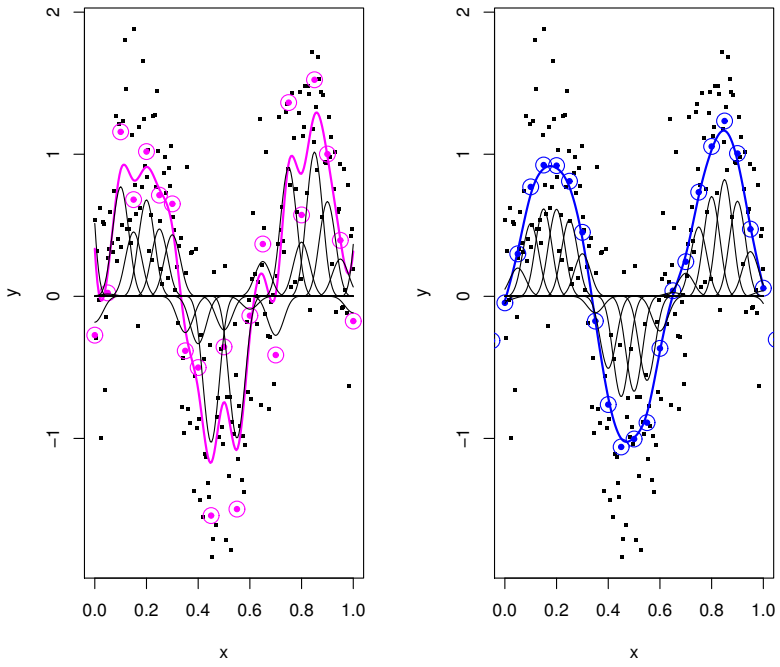
donde $B_m(x)$ son las funciones bases B-splines.

El tipo de penalización está estrechamente ligado al tipo de base que se utilice. Si se utilizan polinomios truncados, la penalización es "*ridge*" independientemente del grado de los polinomio truncados. Para las bases B-splines el término de penalización frecuentemente utilizada es la integral de la segunda derivada al cuadrado de la curva (función B-spline), tal como fue sugerida por O'Sullivan (1986),

$$(1/2)\lambda \int [f''(x)]^2 dx. \tag{16}$$

Supongamos que tenemos una base B-splines construida con $k$ nodos y utilizamos mínimos cuadrados penalizados para ajustar el modelo.

La Gráfica 4 muestra el ajuste de una curva mediante bases B-splines sin y con penalización, las funciones que forman la base multiplicadas por los coeficientes, así como los coeficientes (en círculo). Como se puede observar, el efecto de la penalización fuerza a los coeficientes a seguir un patrón más suave. En la parte izquierda se aprecia como el patrón errático de los coeficientes da lugar a una curva poco suave, en cambio en la parte derecha, cuando se impone que se pase de un coeficiente a otro de forma suave, la curva también lo es.




Gráfica 4. Curva estimada con 20 nodos mediante las bases B-splines sin (izquierda) y con penalización (derecha).

# Capítulo 4
# Modelo de regresión de Cox con splines

En este capítulo se presenta una breve revisión de literatura y se describe la base teórica del modelo de regresión de Cox utilizando los splines, particularmente del modelo de Cox con regresión splines y modelo de Cox con P-splines. Así mismo, se describe los procedimientos de diagnóstico del modelo.

## 4.1. Revisión de la literatura

En muchos trabajos de investigación médica es muy común que en cada paciente además de ser observado el tiempo de supervivencia sean observadas las características clínicas de los pacientes, llamadas covariables. Si el interés es determinar el efecto de las covariables en el tiempo de supervivencia, el propósito del estudio se centra en el análisis de las relaciones entre el tiempo de supervivencia y las covariables mediante un modelo de regresión.

Los modelos parametricos son eficientes cuando se tiene información del modelo de distribución subyacente a las variables y sólo resta por determinar un número finito de parámetros; sin embargo, una fuente de error puede ser elegir una familia paramétrica no adecuada. En estos casos podemos utilizar los modelos semi-paramétricos o no-paramétricos que además de permitir graduar las probabilidades brutas que no siguen un modelo paramétrico establecido, pueden utilizarse para proporcionar una prueba de diagnóstico de los modelos paramétricos o simplemente para explorar los datos.

En análisis de datos de supervivencia, el modelo de regresión semiparamétrica muy frecuentemente utilizado es el modelo de riesgos proporcionales de Cox (Cox, 1972), llamado generalmente como modelo de Cox. Este modelo asume que la función de riesgo es constante sobre un periodo de tiempo y el efecto de las covariables se relaciona linealmente con el logaritmo de la razón de riesgos. Si los supuestos del modelo no se cumplen, el modelo de Cox no es el más adecuado, entonces el modelo de Cox estratificado o el modelo de Cox con variables tiempo-dependiente podrían ser una alternativa. Otras alternativas pueden ser el modelo de odds proporcional y el modelo log-logístico.

Sin embargo, el modelo de Cox estratificado no es adecuado cuando si se tiene más de una variable que no verifica el supuesto de riesgos proporcionales, ya que se pierde la información para estas variables que pueden ser de importancia para quién investiga. En el modelo de Cox con variable tiempo-dependiente, así como en los demás modelos alternativos si el efecto de las covariables presentará una forma funcional no-lineal, el problema de modelado aún estaría pendiente.

Durante las dos últimas décadas numerosas técnicas se han desarrollado para explorar la forma funcional del efecto de las covariables de una manera más flexible, utilizando para ello métodos de suavizamiento (smoothing spline) y polinomios fraccionales (fractional polynomials). En este trabajo se utiliza los splines penalizados (penalized splines) como una alternativa en situaciones en que la forma funcional del efecto de las covariables en la función de riesgo es no-lineal.

Los métodos splines son una herramienta usual en muchos contextos estadísticos, permiten el manejo de relaciones no lineales complejas, difíciles de calcular con modelos paramétricos convencionales. Los splines más utilizados son los splines penalizados (P-splines). Los P-splines (Eilers y Marx, 1996) aproximan una función desconocida por un spline polinomial que puede ser escrito como una combinación lineal de funciones bases splines.

Las diferentes aproximaciones mediante splines son bastante amplias, abarcando desde técnicas de suavización splines (Hastie y Tibshirani (1990), Wahba (1990), Green y Silverman (1994)) con nodos fijos, hasta el uso de regresión splines con selección adaptativa de los nodos (Friedman, 1999). Eilers y Marx (1996) propusieron el uso de splines penalizado (P-splines), un enfoque diferente que puede ser visto como un compromiso entre suavizamiento spline y regresión spline. En éste, el número de nodos que define la función spline es mayor que el justificado, pero menor al número de observaciones. El nivel de sobre-ajuste (overfitting) es controlado por una "roughness penalty" sobre la curva.

En el modelo de Cox, O'Sullivan (1988) utiliza splines de suavizado para estimar el efecto no-lineal de la covariable. Sleeper y Harrington (1990) utilizan regresión splines con un número reducido de nodos, donde los coeficientes son estimados utilizando el método estándar sin funciones de penalización; sin embargo, son más sensibles al número y localización de los nodos y por tanto, son más inestables. Gray (1992) utiliza splines penalizados (penalizad splines) en el modelo de Cox.

Para los propósitos de este trabajo, aquí se presentan las bases metodológicas de los modelos de regresión de Cox utilizando los splines para determinar los factores pronósticos para la supervivencia de los pacientes con LNH y se estima la forma funcional del efecto de las covariables en la razón de riesgo (hazard ratio).

En esta sección, previamente se hace una breve descripción del modelo de Cox clásico, luedo se desarrolla el modelo de Cox con regresión splines y el modelo de Cox con suavizamiento splines penalizado (P-spline), y después se describe los métodos para la verificación de los supuestos del modelo de Cox en general. Lo atractivo del modelo de Cox con splines, es que este modelo permite describir la forma funcional no-lineal de los efectos de las covariables en la función de riesgo en lugar de simplemente determinar el efecto significativo de las covariables.

## 4.2. Modelo de regresión de Cox

Sean los datos observados en la muestra de la forma ($T$,$\delta$,$X$), donde, $T$ es el tiempo de supervivencia observada, $\delta$ el indicador de censura y $X$ las covariables. Sea el objetivo del análisis evaluar el efecto de las covariables en el tiempo de supervivencia hasta la ocurrencia de un evento de interés (ejemplo: falla, muerte, recurrencia u otros). En esta situación el modelo de regresión de Cox es una alternativa para analizar el efecto de las covariables en el tiempo de supervivencia.

### 4.2.1. Modelo de Cox clásico.

El modelo de riesgos proporcionales introducido por Cox (1972), llamado modelo de regresión de Cox o modelo de Cox, es de la forma

$$\lambda(t/X) = \lambda_0(t)\exp\{\mathbf{X}'\beta\}, \tag{17}$$

donde $\lambda_0(t)$ es la función de riesgo basal cuando $X = 0$ (cuya distribución es no especificada), $\beta = (\beta_1, \ldots, \beta_p)$ es un vector de parámetros del modelo y $X = (X_1, \ldots, X_p)$ son los vectores de covariables.

En el modelo de Cox, la función de riesgo es el producto de la función de riesgo basal $\lambda_0(t)$ y un escalar $\exp\{\mathbf{X}'\beta\}$ que sólo depende de los parámetros y las covariables. Las cuales, imponen las siguientes restricciones al modelo (17):

- la razón entre las funciones de riesgo para dos individuos es proporcional (*riesgo proporcional*)
- el logaritmo de la razón de riesgo es independiente del tiempo (*riesgo constante*)
- el logaritmo de la razón de riesgo se relaciona linealmente con las covariables ($forma\ funcional\ lineal$).

Es decir, para dos individuos con valores $u$ y $v$ la relación

$$\lambda(t/u)/\lambda(t/v) = \exp(\beta_j(u_1 - v_1) + \ldots + \beta_p(u_p - v_p))$$

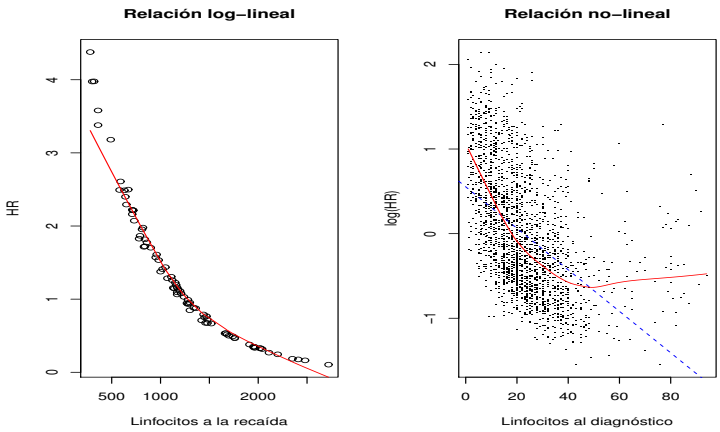
no implica $\lambda_0(t)$ (de ahí el nombre de modelo de riesgos proporcionales); condición que debe ser verificado para las variables categóricas.

La relación

$$\lambda(t/X)/\lambda_0(t) = \exp(\beta_1 x_1 + \ldots + \beta_p x_p)$$

no depende del tiempo, lo que implica que el modelo es de riesgo constante; condición que debe ser verificado para las variables contínuas.

La gráfica 5, muestra un ejemplo de cumplimiento o no de la relación lineal entre el logaritmo de la razón de riesgo (hazard ratio) y las covariables. En la gráfica se observa que el efecto de los linfocitos a la recaída de los pacientes con cáncer de mama (gráfica de la derecha) si cumple el supuesto de relación lineal, en cambio los linfocitos al diagnóstico de los pacientes con LNH no verifica dicho supuesto (gráfica de la izquierda).




GRÁFICA 5. Forma funcional lineal y no-lineal del efecto de los linfocitos.

En el modelo de la forma (17), el problema de modelamiento de los datos de supervivencia se reduce a la estimación de los parámetros $\beta$ y $\lambda_0(t)$ condicionada a $\hat{\beta}$, y en realizar la prueba de hipótesis sobre los parámetros, $H_0 : \beta = \beta_0$, para evaluar el efecto de las covariables en la función de riesgo.

### 4.2.2. Estimación de los parámetros.

Sea una muestra de tamaño $n$, donde los datos observados en la muestra son las ternas $(T_i, \delta_i, X_i)$. Asumimos que los tiempos de supervivencia son censurados por la derecha bajo un mecanismo de censura no-informativa.

En el modelo de la forma (17) la estimación de los parámetros $\beta$ deben ser estimados a partir de las observaciones muestrales. La presencia de la componente no-paramétrica invalida el uso del método de máxima verosimilitud. Para estimar el vector de parámetros $\beta$, Cox (1972) propone el método de verosimilitud parcial.

Supongamos que los datos están compuestos por $n$ individuos, de los cuales existen $r$ tiempos de muertes diferentes, los restantes $n - r$ tiempos se consideran censurados. Así mismo, se supondrá que solo un individuo muere en cada tiempo, es decir no hay empates.

Sea $t_{(1)} < \ldots < t_{(r)}$ los distintos tiempos de muerte ordenados $(r \leq n)$, $R(t_{(j)}) = R_j$ el conjunto de individuos que se encuentran a riesgo en el tiempo $t_{(j)}$, es decir, conjunto de individuos que se encuentran con vida y sin censura antes de $t_{(j)}$, y sea $X_{(j)}$ el vector de covariables asociadas.

Cox (1972) define que la función de verosimilitud parcial para el modelo de riesgos proporcionales como

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(X'_{(j)}\beta)}{\sum\limits_{l \in R_j} \exp(X'_l \beta)} \tag{18}$$

y su logaritmo puede expresarse se como

$$\ell(\beta) = ln(L(\beta)) = \sum_{j=1}^{r} \left\{ X'_{(j)}\beta - ln \sum_{l \in R_j} \exp(X'_l \beta) \right\} \tag{19}$$

Equivalentemente incluyéndose toda la muestra, sea $t_{(1)} < \ldots < t_{(n)}$ las distintas observaciones de tiempos ordenados, $\delta_{(i)}$ indicador de censura respectiva y $X_{(i)}$ el vector de covariables asociadas.

Sea $m_{(i)}$ el evento (muerte) de un individuo $i$ en el instante $t_{(i)}$ y sea $R(t_{(i)}) = R_{(i)}$ el conjunto de individuos en riesgo en el tiempo $t_{(i)}$. Dada la función de riesgo de la forma (17) para cada sujeto $i$ y definida la probabilidad de observar una falla en un individuo $i$ de la forma

$$P(m_{(i)}/t_{(i)}\epsilon R(t_{(i)})) = \left\{ \exp(X'_{(i)}\beta) \Big/ \sum_{l \in R(t_{(i)})} \exp(X'_l\beta) \right\}^{\delta_{(i)}},$$

la función de verosimilitud parcial es dado por

$$Ł(\beta) = \prod_{i=1}^{n} \left\{ \exp(X'_{(i)}\beta) \Big/ \sum_{l \in R(t_{(i)})} \exp(X'_l\beta) \right\}^{\delta_{(i)}}, \tag{20}$$

y el logaritmo de la función de verosimilitud parcial es dado por

$$\ell(\beta) = ln[L(\beta)] = \sum_{i=1}^{n} \delta_{(i)} \left\{ X'_{(i)}\beta - ln \sum_{l \in R(t_{(i)})} \exp(X'_l\beta) \right\}. \tag{21}$$

Cuando los datos presentan tiempos observados empatados, la función de verosimilitud parcial (18) es modificada de alguna forma. Se han propuesto varias aproximaciones para la función de verosimilitud parcial en esta situación, por ejemplo, Breslow (1974), Efron (1977) y Cox (1972).

Sea $t_{(1)} < \ldots < t_{(r)}$ los $r$ distintos tiempos observados y ordenados. Sea $d_j$ el número de fallos observadas en $t_{(j)}$ y $D_j \equiv D(t_{(j)}) = j_1, \ldots, j_{d_j}$ el conjunto de etiquetas de los individuos que fallan en $t_j$. Sea $S_j = \sum_{l \in D_j} X_l$ y $R_j$ el conjunto de subíndices de individuos que se encuentran en riesgo antes de $t_j$.

La aproximación de la verosimilitud parcial sugerida por Breslow (1974) considera que las $d_j$ fallas al tiempo $t_{(j)}$ son distintos y ocurren secuencialmente. Cuando se tienen pocos empates, esta proporciona una muy buena aproximación de la función de verosimilitud parcial.

La verosimilitud debida a Breslow (1974), en el caso de empates, es

$$\prod_{j=1}^{r} \frac{\exp(S'_j\beta)}{\left[ \sum_{l \in R_j} \exp(X'_l\beta) \right]^{d_j}} \tag{22}$$

Una aproximación sugerida por Efron (1977) es

$$\prod_{j=1}^{r} \frac{\exp(S'_j\beta)}{\prod_{k=1}^{d_j} \left[ \sum_{l \in R_j} \exp(X'_l\beta) - (k-1)d_k^{-1} \sum_{l \in D_j} \exp(X'_l\beta) \right]} \tag{23}$$

Otra aproximación sugerida por Cox (1972) es

$$\prod_{j=1}^{r} \frac{\exp(S_j'\beta)}{\sum\limits_{l \in R(t_{(j)};d_j)} \exp(S_l^{*'}\beta)} \tag{24}$$

donde $R(t_{(j)}; d_j)$ denota el conjunto de todos los subconjuntos de $d_j$ individuos seleccionados del conjunto de riesgo $R(t_j)$ sin reemplazo. De este modo, si $\ell \in R(t_{(j)}; d_j)$, éste es de la forma $\ell_1, \ldots, \ell_{d_j}$. La verosimilitud parcial anterior es computacionalmente difícil si el número de empates es grande.

A partir de cualquiera de las expresiones (18), (21), (22), (23) y (24), se pueden obtener los estimadores de máxima verosimilitud parcial de los parámetros $\beta = (\beta_1, \ldots, \beta_p)$ y se pueden realizar las pruebas de hipótesis basadas en la distribución asintótica de los estimadores.

El estimador de máxima verosimilitud del vector de parámetros se obtiene como una solución del sistema de $p$ ecuaciones no lineales generadas por las derivadas parciales de $\ell(\beta)$ respecto a los parámetros, $\beta_j$, $d\ell(\beta)/d\beta_j = 0$ $(j = 1, \ldots, p)$ y utilizando procedimientos iterativos como el método de Newton-Raphson.

### 4.2.3. Prueba de hipótesis e inferencia.

Bajo las condiciones de regularidad (consistencia, distribución asintótica normal, eficiencia asintótica), se dice que los estimadores de máxima verosimilitud tienen distribución asintóticamente normal con media $\beta = (\beta_1, \ldots, \beta_p)$ y matriz de varianza y covarianza $I^*(\beta)$. Dada las complicaciones en el cálculo de $I^*(\beta)$ es común utilizar en su reemplazo la matriz de información observada $I(\beta)$.

Sea la función de puntaje definida como la derivada parcial del logaritmo de la función de verosimilitud con respecto a los parámetros, $U_j(\beta) = d\ell(\beta)/d\beta_j$ $(j = 1, \ldots, p)$ y la matriz de información como el negativo de la derivada del puntaje de eficiencia con respecto a los parámetros, $I(\beta) = [-dU(\beta)/d\beta_k]$ $(k = 1, \ldots, p)$.

En el modelo de la forma (17) la prueba de hipótesis global, $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$, para una muestra de tamaño $n$ suficientemente grande, bajo la distribución asintótica de los estimadores de máxima verosimilitud parcial, se pueden realizar mediante tres estadísticos de prueba diferentes: la prueba de Wald, la razón de verosimilitud y la prueba del score, que bajo la hipótesis nula todas ellas tienen distribución $\chi^2$ con $p$ grados de libertad.

**Prueba de hipótesis sobre el efecto de los parámetros del modelo**:

Sean $b = (b_1, \ldots, b_p)$ los estimadores de máxima verosimilitud del vector de parámetros desconocidos, $\beta = (\beta_1, \ldots, \beta_p)$ y $I(\beta)$ la matriz de información. Los estadísticos de prueba para contrastar la hipótesis, $H_0 : \beta = \beta_0$, basados en la distribución asintótica de los estimadores son:

- La prueba basada en la normalidad asintótica de los estimadores, referida como la prueba Wald: $\chi^2_W = (b - \beta_0)'(I(b))(b - \beta_0) \sim \chi^2_p$
- La prueba de razón de verosimilitud: $\chi^2_{LR} = 2[\ell(b) - \ell(\beta_0)] \sim \chi^2_p$
- La prueba de score: esta prueba es basada en el puntaje de eficiencia $U(\beta) = (U_1(\beta), \ldots, U_1(\beta))$. Para muestra grande, $U(\beta)$ es asintóticamente normal $p$-variada con media 0 y matriz de varianza y covarianza $I(\beta)$, $\chi^2_{Sc} = U(\beta_0)'\ I^{-1}(\beta_0)U(\beta_0) \sim \chi^2_p$

**Prueba de hipótesis sobre el efecto individual de los parámetros del modelo**:

Para contrastar la hipótesis $H_0 : \beta_j = 0$ versus $H_0 : \beta_j \neq 0$, se puede utilizar el estadístico $\hat{\beta}_j / \sqrt{\hat{v}(\hat{\beta}_j)}$, donde $v$ es la varianza del estimador del parámetro $\beta_j$.

Finalmente, de los resultados de modelo de Cox (17), para los propósitos de interpretación de los resultados en términos de riesgo, la razón de riesgo (hazard ratio) para una covariable binaria se puede estimars mediante $\hat{H}R = \exp\{\widehat{\beta}\}$ y para una covariable contínua dada $(x)$ mediante $\widehat{H}R = \exp\{\hat{\beta}x\}$ y la curva de supervivencia mediante $\hat{S}(t/X) = \exp(-\hat{H}(t/X)) = \exp\{-\int_t^\infty \hat{\lambda}(s/X)ds\}$ versus $t$.

Por otro lado, debido a que los métodos de diagnóstico del modelo están desarrollados en el contexto de proceso de conteo y teoría de martingala, en los párrafos siguientes se describe el modelo de Cox en el contexto de proceso de conteo.

### 4.2.4. Modelo de Cox en el contexto de procesos de conteo.

El tratamiento de datos de supervivencia mediante procesos de conteo tiene sus orígenes en el trabajo de Aalen (1978). Posteriormente Andersen y Gill (1982) integraron el modelo de Cox en el marco de procesos de conteo, generalizando de esta forma el tratamiento habitual de los modelos de supervivencia. Andersen y Gill (1982) extienden el modelo Cox en el contexto de procesos de conteo y obtienen pruebas martingala para las propiedades asintóticas de los estimadores asociados (Martinussen and Scheike, 2006).

Supongamos que observamos $n$ observaciones de la forma $(T_i, \delta_i, X_i)$, donde $T_i$ es el tiempo de supervivencia censurado por la derecha, $\delta_i$ en indicador de censura, $X_i$ el vector de covariables. El modelo de Cox, asume que la función de intensidad es de la forma

$$\lambda(t) = Y(t)\lambda_0(t)\exp(X'\beta), \tag{25}$$

donde $Y(t)$ es indicador de riesgo que es uno si el evento no ha ocurrido, $\lambda_0(t)$ es la función de riesgo base no-paramétrico localmente integrable, $X = (X_1, \ldots, X_p)$ es un vector de $p$ covariables y $\beta$ es el vector de parámetros del modelo.

Sea una muestra de tamaño $n$ conteniendo datos de la forma $(N_i(t), Y_i(t), X_i)$ $i = 1, \ldots, n$ que son observados en un intervalo de tiempo $[0, t]$, $t < \infty$, y que cada $N_i(t)$ tiene intensidad de la forma (25).

Los parámetros $\beta$ del modelo son estimados maximizando como en la función de verosimilitud parcial de Cox (Cox (1972), Martinussen y Scheike (2006)),

$$L(\beta) = \prod_t \prod_i \left( \frac{\exp(X_i'\beta)}{S_0(t,\beta)} \right)^{\Delta N_i(t)},$$

donde $S_0(t,\beta) = \sum_{i=1}^n Y_i(t)\exp(X_i'\beta)$.

El estimador $\hat{\beta}$ se obtiene como una solución de la ecuación del score $U(\beta) = 0$, donde $U(\beta) = \sum_{i=1}^n \int_0^\tau (X_i - \frac{S_1(t,\beta)}{S_0(t,\beta)} dN_i(t)$ y $S_1(t,\beta)$ es la derivada parcial de primer orden de $S_0(t,\beta)$ con respecto a $\beta$.

Si $\beta$ es fijado, el estimador de Nelson-Aalen de $\Lambda_0(t)$ es estimado como

$$\hat{\Lambda_0}(t,\beta) = \int_0^t \frac{1}{S_0(s,\hat{\beta})} dN(s),$$

donde $N(t) = \sum_t N_i(t)$. Así mismo, dado $\hat{\beta}$ como una solución de $U(\beta) = 0$, el estimador de Breslow de $\Lambda_0(t)$ es dado por $\hat{\Lambda}_0(t) = \hat{\Lambda}_0(t,\hat{\beta})$.

Bajo la distribución asintótica de los estimadores de máxima verosimilitud parcial, los estadísticos de prueba son validos para realizar la prueba de hipótesis sobre los parámetros $\beta$. Sea $I(\beta)$ el negativo de la primera derivada de la función score con respecto a $\beta$ y sea $\beta_0$ que denota el verdadero valor de $\beta$.

En consecuencia la prueba de hipótesis sobre la hipótesis, $H_0 : \beta = \beta_0$, se puede realizar mediante los estadísticos de prueba Wald, de razón de verosimilitud o el de score, similar al procedimiento clásico en el modelo de Cox.

## 4.3. Introducción al modelo de Cox con splines

El modelo de riesgos proporcionales de Cox (1972) es un modelo de regresión de uso muy frecuente, para analizar el efecto de las covariables en la supervivencia. En este modelo la respuesta modelada es la función de riesgo, con logaritmo de la razón de riesgo que es lineal en las covariables.

El modelo de Cox clásico (17) es conceptualmente atractivo, porque es una medida de asociación de la razón de riesgo (hazard ratio) que es simplemente una función de los parámetros de regresión. La restricción natural del modelo implica que la forma del riesgo base, no necesita ser especificada para la estimación de los parámetros del modelo, ya que esta no entra en la función de verosimilitud parcial (18).

En este modelo, la razón de riesgo al tiempo $t$ para dos pacientes con valores de las covariables $u_1, \ldots, u_p$ y $v_1, \ldots, v_p$, respectivamente, no implica el riesgo base, y de ahí el nombre de modelo de riesgos proporcionales. La razón de riesgo para dos individuos con los mismos valores para las covariables, pero que reciben diferentes tratamientos, es simplemente $e^{\beta_l}$ (suponiendo que $x_l$ es el indicador de tratamiento). Así mismo, la razón de riesgo no depende en el tiempo, lo que implica que el modelo de la forma (17) es de riesgo constante, y que el logaritmo de la razón de riesgo se relaciona linealmente con el efecto de las covariables.

Sin embargo, en muchas situaciones algunas covariables (como la edad, el estadio de la enfermedad, etc.) puede ser relacionados linealmente con la supervivencia, mientras otras como los parámetros de laboratorio pueden ser descritas con más precisión por una relación no lineal. Ignorar o una especificación inadecuada de estos efectos podrían sesgar los resultados de las estimaciones, por lo que los efectos insignificantes podrían parecer importantes o viceversa.

Los métodos descritos aquí relajan los supuestos de linealidad y permiten sin problemas aproximar la forma funcional del efecto no lineal de las covariables en el logaritmo de la razón de riesgo, utilizando los métodos de suavizamiento basado en las funciones splines, denominado "métodos de suavizamiento mediante splines"(descrito en la sección 3.5). Una relación lineal será un caso particular.

Las diferentes aproximaciones mediante splines son bastante amplias, abarcando desde técnicas de suavización splines con nodos fijos (Hastie y Tibshirani (1990), Wahba (1990), Green y Silverman (1994)), hasta el uso de regresión splines con selección adaptativa de los nodos (Friedman, 1999).

En el modelo de Cox, O'Sullivan (1988) utiliza suavizamiento splines, Sleeper y Harrington (1990) regresión splines con un número reducido de nodos (aunque estas son más sensibles

al número y localización de los nodos y por tanto, son más inestables). Gray (1992) utiliza suavizamiento splines penalizados (P-splines); el cual, es un término intermedio entre suavizamiento splines y regresión splines.

En esta sección se describe a manera de introducción el modelo de Cox con componentes no lineales, los métodos de suavizamiento mediante los splines y se plantea el modelo de riesgos proporcionales general (modelo aditivo parcialmente lineal) que incluye componentes lineales y no lineales sobre las que se desarrollará las secciónes 4.4 y 4.5 y se realizarán las aplicaciones.

### 4.3.1. Modelo de Cox con componentes no lineales.

En esta subsección se considera la notación usual de los datos de supervivencia. Los datos evaluables son de la forma ($T$,$\delta$,$X$), donde $T$ es el tiempo de supervivencia, $\delta$ es el indicador de censura y $X$ el vector de covariables.

El modelo de Cox con $p$ covariables no lineales (componentes no lineales) puede ser expresado de la forma:

$$\lambda(t|X) = \lambda_0(t)\exp\{\sum_{j=1}^{p} s_j(x_j)\}, \tag{26}$$

donde las $x_j$ son las covariables y $s_j(x_j)$ son las funciones que aproximan el efecto de las covariables de una manera suave en el logaritmo de la función de riesgo.

El modelo (26) es conocido como "modelo aditivo generalizado" (Hastie y Tibshirani, 1990). El efecto de cada covariable en el logaritmo de riesgo es aditivo y es representado por una suave, función posiblemente no-lineal. Las transformaciones de $s_j$ no son elegidos a priori (por ejemplo, logaritmo o exponencial), sino que se estima de manera flexible a partir de los datos.

En casos raros, todas las transformaciones de las covariables pueden ser funciones no-lineales suavizadas. Por lo general algunas covariables (incluyendo la variable de tratamiento) son factores; estos se modelan como funciones de paso con una constante que representa cada nivel, o algunos contrastes predifinidos de los niveles. Otras covariables cuantitativas como la edad o el peso puede ser modelado no linealmente, aunque si un ajuste lineal es suficiente, esto es usualmente preferido por su simplicidad.

Una característica atractiva del modelo aditivo generalizado, en particular en el modelo de la forma (26), es que este modelo disminuye la necesidad de categorizar una covariable continua con el fin de descubrir la naturaleza del efecto de las covariables en la función

de riesgo; particualmente la forma funcional del efecto de las covariables no lineales en el logaritmo de la razón de riesgo (hazard ratio).

En general, en el modelo de la forma (26), exisen dos formas de ajustar los términos no-lineales $s_j(x_j)$ en el modelo de riesgos proporcionales. El primer método requiere que se seleccione la forma y la complejidad de la función que mejor aproxime $s_j(x_j)$; lo cual es fácil de implementar y el modelo resultante se parece al modelo de Cox clásico. El otro método es más automático, es decir, "basado en datos" con una ligera sobrecarga computacional.

En la subsección seguiente, se describe de manera breve los métodos de aproximacion de la forma funcional del efecto de las covariables, $s_j(x_j)$, mediante los métodos de suaviamiento basado en los splines, paritularmente el modelo de Cox con regressión splines y modelo de Cox con suavizamiento splines penalizado, que se desarrollará con más detalle en las secciones 4.4 y 4.5.

### 4.3.2. Métodos de suavizamiento en el modelo de Cox.

En esta subsección, nos limitaremos al caso de un modelo con una sola covariable $x$ y la estimación de la función $s$, mediante métodos de suavizamiento con splines (regresión splines y P-splines). Los métodos descritos se puede extender al caso de covariables múltiples de una manera similar (Hastie y Tibshirani, 1990).

#### *4.3.2.1. Modelo con regresión splines.*

Las funciones polinomiales son a menudo utilizadas para modelar las transformaciones no-lineales. La flexibilidad aumenta con el grado del polinomio, resultando en una aproximación razonable para la función $s$ en muchos casos. Sin embargo, el ajuste de un polinomio para $s$ puede ser influenciado por valores extremos.

Polinomios por trozos actuan localmente en las regiones definidas por los nodos y proporcionan una aproximación más estable, ya que cada trozo es generalmente de menor grado que el requerido por un polinomio simple. Un spline polinomial de grado $d$ es una función polinomial a trozos que tiene $d-1$ derivadas continuas en cada nodo; para $d \geq 3$, la función es visualmente suave. La función spline es expresada como una combinación lineal de los elementos bases.

Aunque hay un amplia discusión sobre los splines y los polinomios a trozos, nuestra revisión se centra en los splines de bajo grado con uno o más nodos interiores, en particular, los splines cúbicos. El espacio de todos los splines cúbicos con una secuencia determinada de $k$ nodos y las correspondientes condiciones de continuidad pueden ser generadas a partir de un conjunto de funciones bases.

Modelo basado en bases de potencia truncada:

Considerando primero el espacio de un simple polinomio cúbico; una base es dada por $1, x, x^2$ y $x^3$, y cualquier función polinomial cúbica en $x$ puede ser escrito como $\alpha_0 + \alpha_1 x_1 + \alpha_2 x^2 + \alpha_3 x^3$. Considerando un simple nodo en $x = \xi_1$ e incluyendose la función base $(x - \xi_1)^3_+$ en el polinomio cúbico, no es defícil verificar que

$$s(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 (x - \xi_1)^3_+ \tag{27}$$

satisface las condiciones para un spline cúbico dado el nodo. La función (27), es conocida como función splines con bases de potencia truncada. Para un spline polinomial de grado $d$ y $k$ nodos, el espacio es de dimensión $d + k + 1$.

En consecuencia, el modelo de Cox con regresión splines, basado en bases de pontencia truncada de grado tres, se puede expresar como,

$$\lambda(t|X) = \lambda_0(t) \exp\{s(x)\} = \lambda_0(t) \exp\{\alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 (x - \xi_1)^3_+\}, \tag{28}$$

donde, $\alpha_0$ es abosrbido por la función de riesgo base.

El modelo (28) es similar al modelo de Cox con 4 parámetros, por tanto la estimación de los parámetros se puede realizar utilizando la función de máxima verosimilitud parcial, similar para el caso del modelo de Cox clásico.

Modelo basado en bases B-splines:

Una forma menos intuitiva pero atractiva computacionalmente base para la representación de los splines son las bases B-splines. Cada función base B-spline de grado $d$ es distinto de cero y este comportamiento local es explotado en los cálculos cuando el número de nodos es muy grande (deBoor, 1978).

Una función spline $s(x)$ es expresado como una combinación lineal de estas funciones bases B-splines (descrito en la sección 3.5):

$$s(x) = \sum_{l=1}^{k+d+1} \alpha_l B_l(x) \tag{29}$$

donde $d$ es el grado del spline, $k$ el número de nodos, y $B_l(x)$, $l = l, \ldots, k + d + 1$ son los elementos bases B-spline de grado $d$ con $k$ nodos.

En consecuencia, el modelo de Cox con regresión splines basado en bases B-splines, se puede expresar como

$$\lambda(t|\mathbf{X}=\mathbf{x}) = \lambda_0(t)\exp\left\{\sum_{l=1}^{d+k}\alpha_l B_l(x)\right\} \tag{30}$$

donde, el término constante es absorbida en la función de riesgo base.

Note que la forma para $s(x)$ se parece al predictor lineal del modelo de Cox. La covariable $x$ se expande en un conjunto de "nuevas variables", y los coeficientes $\alpha_1, \ldots, \alpha_p$ pueden ser estimados utilizando el método de máxima verosimilitud parcial y se puede realizar las pruebas de hipótesis similar al caso del modelo de Cox clásico, si asumimos que el número y la localización de los nodos es fijo.

El uso de regresión splines para identificar las posibles tendencias no lineales de las covariables continuas (o al menos ordinales) es atractiva porque no necesita de un software especial, una vez que los B-splines son evaluados en cada valor observado.

El inconveniente de la regresión splines es que el analista debe elegir el número y la localización de los nodos. A menos que haya razones para colocar los nodos en los puntos concretos, una estrategia razonable es colocar en los cuantiles del predictor en cuestión; es decir, un solo nodo en la mediana, dos nodos en las terciles, etc.

En aplicaciones con datos de los ensayos clínicos, se recomienda no más de cuatro nodos para describir con precisión la relación subyacente entre las covariables y el endpoint de la enfermedad, en particular para datos de supervivencia.

Una aplicación detallada utilizando bases B-splines en el modelo de Cox con regresión splines para identificar los factores pronósticos para la enfermedad hepática se da en Sleeper y Harrington (1990).

En el package R, la función *coxph* permite realizar el ajuste del modelo de Cox con regresión splines, por ejemplo con bases B-splines (función *bs*).

*4.3.2.2. Modelo con suavizamiento splines.*

El método de suavizamiento splines son muy usuales en muchas aplicaciones, porque aproximan bien las funciones conocidas y funciones no-lineales. A diferencia de los splines de la sección anterior, donde la suavidad es impuesta por el número de nodos, en suavizamiento splines basado en bases B-splines se tiene un nodo en cada valor de $x$. El grado de suavidad del splines resultante $s$ es controlado usando una función de penalización de la rugosidad (Hastie y Tibshirani, 1990).

Sea el modelo de Cox con un término no lineal, expresado de la forma

$$\lambda(t|\mathbf{X} = \mathbf{x}) = \lambda_0(t) \exp\{g(x)\} \tag{31}$$

donde $g(x)$ es una función que transforma la covariable $x$, que en este caso es una función splines $g(x) = s(x)$ que aproxima suavemente la forma funcional no lineal de la covariable $x$.

En el modelo de la forma (31), como son discutidos en Sleeper y Harrington (1990) y Gray (1992), la función splines $s(x)$ es una aproximación natural para la transformación de la covariable $x$, $s(x)$. Más precisamente, sean $B_1(x), \ldots, B_{k+4}(x)$ las bases B-spline cúbico para el espacio de los splines cúbicos con $k$ nodos preestablecidos. Dado que el interés está en las alternativas que se apartan de la expresión lineal de forma suave, un spline cúbico con bases B-splines se puede expresar como,

$$s(x) = \alpha_0 x + \sum_{l=1}^{k+2} \alpha_l B_l(x) \tag{32}$$

donde $\alpha_l$ son los coeficientes de las bases B-splines y $B_l$ son las elementos de las bases B-splines, en este caso para un splines cúbico.

En la expresión (32) sólo se incluyen $k + 2$ (de la $k + 4$) bases splines debido a que el término constante es absorbido en la función de riesgo base y el término lineal se especifica por separado.

Sleeper y Harrington (1990) utilizan la función de verosimilitud parcial para estimar los parámetros, en cambio Gray (1992) resta el término de penalización $\lambda \int [g''(z)]^2 dz$ ($\lambda$ veces la curvatura integral de $g$). Es decir, la estimación se basa en la función de verosimilitud parcial penalizado.

El logaritmo de la función de verosimilitud parcial penalizado es expresado como,

$$\ell_p(\alpha, s; \lambda) = \ell(\alpha, s) - \lambda \int [s(u)''']^2 du, \tag{33}$$

donde $\ell(\alpha, s)$ es el logaritmo de la función de verosimilitud parcial del modelo (31) parametrizado por (32) y $\lambda$ es el parámetro de suavizamiento que controla el grado de suavidad de la curva.

Si $\lambda$ es pequeño, entonces el componente de log-verosimilitud parcial domina y la curva $s$ tiende a seguir los datos de cerca y ser irregular. Si $\lambda$ es grande, el término de penalización domina y la curvas es muy suave $s$. Cuando $\lambda \to \infty$, $s$ es lineal.

En suavizamiento splines penalizado la idea es disminuir la verosimilitud restando el término que da cuenta de la rugosidad de la función $g$. El factor adicional $\lambda$ es un parámetro de ajuste que impone la penalidad; lo cual, gobierna el "trade-off" entre el término de la función de verosimilitud y el término de penalización.

Para un valor fijo del parámetro de suavización, $\lambda$, la verosimilitud parcial penalizado puede ser maximizada para obtener estimaciones de los parámetros. El parámetro de suavizado $\lambda$ puede ser seleccionado por el analista, pero es más apropiado recurrir a métodos que selección automática del valor del parámetro de suavizado, como la validación cruzada (Verweij y van Houwelingen, 1993) o la minimización del criterio de información Akaike (AIC ) (Akaike, 1973).

La estimación no procede hasta que un valor entre cero y el infinito es elegido para $\lambda$. Elegir un simple valor de $\lambda$ es un procedimiento menos subjetivo que elegir el número y la localización de los nodos para una regresión splines. Existe una extensa literatura sobre los métodos automáticos para la elección de $\lambda$, como las técnicas de validación cruzada (Eubank, 1988). Estos son difíciles de aplicar en el contexto actual, y se recurre a un enfoque más subjetivo.

Las unidades de $\lambda$ no son físicamente significativa, y dependen de la escala de medición del factor de pronóstico. Una medida más intuitivo es el equivalente de grados de libertad, $df(\lambda)$, que es definido como $tr(S_\lambda)$; aquí, $S_\lambda$ es la matriz suavizado o operador lineal que produce un ajuste suavizamiento spline, y depende de los valores de $x$ y $\lambda$ (Hastie y Tibshirani, 1990).

Especificado un valor para $df(\lambda)$, se deriva el valor correspondiente para $\lambda$. Por ejemplo, se podría especificar cuatro grados de libertad para un término; esto es más o menos como un polinomio de cuarto orden (en términos de complejidad) y corresponde a una cantidad moderada de suavizado.

En este caso, aunque el suavizamiento splines puede ser especificado menos subjetivamente que la regresión splines, uno paga un precio especial de herramientas computacionales, y los algoritmos son generalmente más lentos. Además, los métodos de inferencia para ajuste con suavizamiento splines, en particular en el caso del modelo de Cox con términos no lineales, son menos sencillos que los utilizados en el ajuste con regresión splines.

Una metodología detallada utilizando el modelo de Cox con suavizamiento splines penalizado con bases B-splines y apliciones para cáncer de mamam se dá en Gray (1992). Gray (1992) utiliza splines penalizados para aproximar las funciones desconocidas, $g$.

Los splines con penalización es un método intermedio entre los métodos de suavizamiento splines y regresión splines; de hecho combina lo mejor de ambos métodos, utilizando menos parámetros que los splines de suavizado, pero la selección de los nodos no es tan importante como en los splines de regresión.

En el package R, la función *coxph* permite realizar el ajuste del modelo de Cox con splines penalizados (P-splines) utilizando bases B-splines (función *psline*).

### 4.3.3. Modelo de riesgos proporcionales general.

Sea el modelo de riesgo proporcional general que incorpora el efecto de las covariables en una forma arbitraria

$$\lambda(t|X) = \lambda_0(t)\exp(g(X)), \tag{34}$$

donde $X$ son las covariables y $g(X)$ es una función que enlaza paramétricamente el efecto de las covariables o es una función no especificada.

En el modelo de la forma ((34)), la función $g(X)$ es asumido para ser una función general que puede ser llamado "covariate function". Esta es una importante e interesante generalización del modelo de Cox. (Dabrowska, 1997; Terje, 1999; Wang, 2009).

Si $X = x$ es una covariable, $g$ puede ser aproximado por una función splines $s$, donde $s$ es expresado como una combinación lineal de las funciones bases splines.

Si $X$ es una matriz de covariables $p$-dimensional, aunque el modelo (34) no restringe el logaritmo de riesgo para ser lineal en $X$, este es una dificultad para estimar $g(X)$ e interpretar el efecto de las covariables en la función de riesgo. El modelo de regresión aditiva (Stone, 1985) facilita una estructura aditiva que permite diferentes funciones para cada covariable. Por lo tanto, en el modelo ((34)), el logaritmo de la razón de riesgo tiene $p$ componentes cada uno representado por una función arbitraria $g_j(x_j)$, $j = 1, \ldots, p$, que puede ser expresado como:

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\{g_1(X_1) + \ldots + g_p(X_p)\}. \tag{35}$$

En el modelo (35) las $p$ funciones desconocidas pueden ser aproximadas mediante splines,

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\{s_1(X_1) + \ldots + s_p(X_p)\}. \tag{36}$$

Los splines son capaces de aproximar las funciones conocidas, por tanto, se espera lo mismo para las funciones componentes en (36).

Considerando las componentes que satisfacen una forma funcional lineal (covariables binarias y otras covariables) y las componentes que no satisfacen la estructura lineal, el modelo (36) puede ser expresado en una forma más general como

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\left\{\sum_{j=1}^{p}\beta_j X_j \ + \sum_{j=p+1}^{p+q} g_j(Z_j)\right\}, \tag{37}$$

donde las primeras $p$ covariables pueden ser covariables binarias o que satisfacen una estructura lineal y las siguientes $q$ covariables no satisfacen la estructura lineal.

En el modelo (37) las $q$ covariables que no satisfacen la forma funcional lineal pueden ser aproximadas mediante métodos de suavizamiento splines. Los métodos de suavizamiento ampliamente reportados en la literatura son: suavizamiento spline y regresión spline, particularmente splines penalizado (P-splines).

En el modelo de Cox para aproximar la función $g$, O'Sullivan (1988) utiliza suavizamiento spline, Sleeper y Harrington (1990) regresión splines, y Gray (1992) splines penalizados.

Utilizando el modelo de la forma más general (37), en las secciones (4.4 y 4.5) se describe con más detalle la metodología del modelo de Cox con regresión splines y suavizamiento splines penalizado (P-splines).

## 4.4. Modelo de regresión de Cox con regresión splines

Sean los datos observados en la muestra de la forma ($T$,$\delta$,$X^*$), donde $T$ es el tiempo de supervivencia observada, $\delta$ el indicador de censura, y $X^* = (X, Z)$ las covariables con $p$ y $q$ componentes con efectos lineales y no lineales, respectivamente; donde el objetivo sea determinar el efecto de las covariables en la supervivencia o riesgo de mortalidad.

En esta sección se describe el modelo de Cox con regresión splines (basado en bases B-splines), como una alternativa para explorar la forma funcional del efecto de las covariables con efectos no lineales en la razón de riesgos (hazard ratio).

### 4.4.1. Modelo de Cox con regresión splines.

El modelo de regresión considerando el modelo de la forma más general (37), que incluye componentes lineales y no lineales (modelo parcialmente lineal), puede ser expresado de la forma,

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp\left\{ \sum_{j=1}^{p} \beta_j x_j \ + \sum_{k=1}^{q} s_k(z_k) \right\}. \tag{38}$$

donde las $x's$ son covariables categóricas o covariables con efectos lineales y las $z's$ son covariables con efectos no lineales que pueden ser aproximados mediante funciones splines $s_k$.

Particularmente, en el modelo (38), cada una de las $q$ componentes no lineales $s_k(z_k)$ pueden ser aproximadas por una combinación lineal de funciones bases B-splines (deBoor, 1978) de grado $d_k$,

$$s_k(z_k) = \sum_{m=1}^{M+d_k} \alpha_{k,m} B_{k,m}(z_k) \tag{39}$$

donde $\alpha_{k,m}$ son los coeficientes de las bases y $B_{k,m}$ los elementos bases B-splines para el $k$-ésimo componente no lineal del modelo.

Una función base splines de grado $d_k$ con $M$ nodos tiene $M + d_k + 1$ funciones, sin embargo en la expresión (39) sólo aparecen $M + d_k$ funciones debido a que el término constante es absorbido en la función riesgo base.

En consecuencia, el modelo de la forma (38) con componentes no lineales parametrizados por (39) tiene $p + d$ ($d = \sum(M + d_k)$) parámetros, similar al modelo de Cox clásico, $[\beta_1, \ldots, \beta_p, \alpha_{1,1}, \ldots, \alpha_{1,M+d_1}, \ldots, \alpha_{q,1}, \ldots, \alpha_{q,M+d_q}]$.

Por tanto, los parámetros se pueden estimar utilizando los procedimientos tradicionales como el método de máxima verosimilitud parcial para el modelo de Cox y realizar las pruebas de hipótesis sobre los parámetros del modelo para determinar el efecto de las covariables en la función de riesgo.

Así mismo, en el modelo (38), se puede estimar el logaritmo de la razón de riesgos mediante $\hat{s}_k(z_k) = \Sigma_{m=1}^{M+d_k} \hat{\alpha}_{k,m} B_k(z_k)$.

### 4.4.2. Estimación de los parámetros.

Sea el predictor aditivo parcialmente lineal del modelo (38) con componentes no lineales parametrizados por (39), denotado como

$$\eta = \left\{ \sum_{j=1}^{p} \beta_j x_j \ + \sum_{k=1}^{q} s_k(z_k) \right\} \tag{40}$$

y los parámetros de la $k$-ésima componente no lineal del modelo como, $\alpha_k = (\alpha_{k,1}, \ldots, \alpha_{k,d_1})$ y al conjunto de parámetros de las componentes lineales y no lineales del modelo $\beta$ y $\alpha$, respectivamente, como $\theta = (\beta, \alpha)$.

Sean los datos observados en una muestra de $n$ individuos; de los cuales existen $r$ tiempos de muerte observados y los restantes $(n - r)$ son considerados como tiempos de supervivencia censurados. Asumiendo que solo un individuo muere en cada tiempo, es decir no hay empates.

Denotando como $t_{(1)} < \ldots < t_{(r)}$ a los distintos tiempos de muerte ordenados ($r \leq n$), $R(t_{(i)}) = R_i$ el conjunto de individuos que se encuentran en riesgo al tiempo $t_{(i)}$, es decir, conjunto de individuos que se encuentran con vida y sin censura antes de $t_{(i)}$, y $X^*_{(i)} = (X_{(i)}, Z_{(i)})$ el vector de covariables asociadas.

El logaritmo de la función de verosimilitud parcial (Cox, 1972) es de la forma,

$$\ell(\theta) = ln(L(\theta)) = \sum_{i=1}^{r} \left\{ \eta_i - ln \sum_{l \in R_i} \exp(\eta_l) \right\} \tag{41}$$

donde $\eta_i$ representa el predictor aditivo del individuo con tiempo de muerte observada y $\eta_l$ el predictor de los individuos en riesgo al tiempo $t_{(i)}$, respectivamente.

El estimador de máxima verosimilitud parcial de los parámetros, $\theta = (\beta, \alpha)$, se obtienen maximizando el logaritmo de la función de verosimilitud parcial (41), como una solución de las ecuaciones generadas por las derivadas parciales del logaritmo de la función de verosimilitud parcial respecto de los parámetros, $\frac{d\ell(\theta)}{d\theta} = 0$, utilizando métodos iterativos como Newton-Rapson.

Sea $U(\theta) = \left(\frac{d\ell(\theta)}{d\theta}\right)$ la función score (puntaje de eficiencia) y $I(\theta) = -\left(\frac{d^2\ell(\theta)}{d\theta d\theta'}\right)$ la matriz de información. Bajo la normalidad asintótica de los estimadores de máxima verosimilitud parcial, los estadísticos de prueba pueden ser construidos exactamente como en la verosimilitud clásica utilizando estas cantidades para realizar las pruebas de hipótesis sobre los parámetros del modelo (38).

Es decir, obtenidos los estimadores máximo verosímiles de los parámetros, así como el puntaje de eficiencia y la matriz de infomación se puede realizar las pruebas de hipótesis sobre los parámetros del modelo aditivo, similar al modelo de Cox clásico, mediante los estadísticos de prueba tipo Wald, de razón de verosimilitud o score.

### 4.4.3. Prueba de hipótesis e inferencia.

En el modelo de la forma (38) existen dos hipótesis de prueba, uno de ellas referidas a los componentes no lineales son de interés particular:

- Hipótesis sobre el efecto global, $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$
- Hipótesis que la $k$-ésima covariable no tiene efecto, $H_{01_k} : \alpha_k = 0$ vs. $H_{01_k} : \alpha_k \neq 0$

Para una muestra, $n$, suficientemente grande, sea $(\hat{\beta}, \hat{\alpha})$ los valores de los parámetros que maximizan el logaritmo de la verosimilitud parcial $\ell(\theta)$. Bajo la distribución asintótica de los estimadores de máxima verosimilitud, se pueden construir los estadísticos de prueba para contraste de hipótesis de los parámetros del modelo (38).

**Prueba de hipótesis sobre el efecto de las componentes del modelo**:

Sea la hipótesis $H_0 : \theta = 0$. Sean $\hat{\theta}$ el estimador de máxima verosimilitud parcial del vector de parámetros, $I(\theta)$ la matriz de información y $a = p + \sum_{k=1}^{q}(M + d_k)$ número de parámetros del modelo. Los estadísticos de prueba para contrastar la hipótesis, basados en la distribución asintótica de los estimadores son dados por:

- Estadístico de prueba tipo Wald: $Q_W = \hat{\theta}'(I(\hat{\theta}))\hat{\theta} \sim \chi^2_a$
- La prueba de razón de verosimilitud: $Q_{LR} = 2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi^2_a$
- Estadístico de prueba tipo score: $Q_{Sc} = U(\theta_0)'\ I^{-1}(\theta_0)U(\theta_0) \sim \chi^2_a$

**Prueba de hipótesis sobre el efecto de las componentes no lineales**:

Sea la hipótesis $H_0 : \alpha = 0$ sobre los parámetros de una componente no lineal $(z_k)$ del modelo (38), donde $\alpha$ es un vector de $c = M + d_k$ parámetros. Sea $\hat{\beta}_0$ el estimador de la máxima verosimilitud parcial para $\beta$ cuando $\alpha = 0$.

Ademas, sea $U(\beta, \alpha) = (U'_\beta(\beta, \alpha), U'_\alpha(\beta, \alpha))$ el score de la verosimilitud parcial e $I$ la matriz de información, con subíndices denotando las submatrices tales como $I_{\alpha\alpha}$. Note que $\left(\frac{d\ell(\hat{\beta}_0,0)}{d\alpha}\right) = U_\alpha(\hat{\beta}_0, 0)$, y que el negativo de la parte $\alpha\alpha$ de la matriz de la segunda derivada del logaritmo de la verosimilitud es $I_{\alpha\alpha}$, con el otro componente los componentes correspondientes de $I$.

Por analogía con el usual procedimiento de verosimilitud paramétrica, los tres diferentes estadísticos de prueba son dados por:

- Estadístico de prueba tipo Wald: $Q_W = \hat{\alpha}'(I_{\alpha\alpha/\beta})\hat{\alpha} \sim \chi^2_c$
- Prueba de razón de verosimilitud: $Q_{LR} = 2[\ell(\hat{\beta}, \hat{\alpha}) - \ell(\hat{\beta}_0, 0)] \sim \chi^2_c$
- Estadístico de prueba tipo score: $Q_{Sc} = U'_\alpha(\hat{\beta}_0, 0)(I_{\alpha\alpha/\beta})^{-1}U_\alpha(\hat{\beta}_0, 0) \sim \chi^2_c$, donde $I_{\alpha\alpha/\beta} = I_{\alpha\alpha} - I_{\alpha\beta}I^{-1}_{\beta\beta}I_{\beta\alpha}$.

**Prueba de hipótesis sobre el efecto individual de las componentes lineales del modelo**:

Para contrastar la hipótesis $H_0 : \beta_j = 0$ versus $H_0 : \beta_j \neq 0$, se puede utilizar la estadística de prueba tipo Wald, $(\hat{\beta}_j/\sqrt{\hat{v}(\hat{\beta}_j)})^2$, donde $v$ es la varianza del estimador del parámetro $\beta_j$. Este esdístico tiene una distribución Chi-cudrado con un grado libertad para una covariable contínua o binaria y $r - 1$ grados de libertad para una covariable con $r$ categorías o niveles.

Aunque este método es más sensible al número y la localización de los nodos que los splines penalizados, y por tanto, son mas inestables, aquí se presenta como parte de una revisión de los mismos.

En el programa R, la función *coxph* con B-splines (*bs*) o natural splines (*ns*) de la librería *survival* y *splines* permiten realizar el ajuste de los datos mediante el modelo de Cox con regresión splines.

## 4.5. Modelo de regresión de Cox con P-splines

Sean los datos observados en la muestra de la forma ($T$,$\delta$,$X^*$), donde $T$ es el tiempo de supervivencia observada, $\delta$ es el indicador de censura, y $X^* = (X, Z)$ las covariables con $p$ y $q$ componentes con efectos lineales y no lineales, respectivamente.

### 4.5.1. Modelo de Cox con P-splines.

Sea el modelo de la forma más general (37), que incluye componentes lineales y no lineales (modelo parcialmente lineal), que puede ser expresado de la forma

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp\left\{ \sum_{j=1}^{p} \beta_j x_j \ + \sum_{k=1}^{q} s_k(z_k) \right\}. \tag{42}$$

donde las $x's$ son las componentes lineales y las $z's$ no lineales. En el modelo (42), cada una de las $q$ componentes no lineales $s_k(z_k)$ puede ser aproximado por una combinación lineal de las funciones bases B-spline (deBoor, 1978).

Sea $s_k(z_k)$ una combinación lineal de bases B-splines cúbico de la forma

$$s_k(z_k) = \theta_{k,0} z_k + \sum_{m=1}^{M+2} \theta_{k,m} B_{k,m}(z_k) \tag{43}$$

Una función base B-spline cúbico con $M$ nodos tiene $M+4$ funciones, debido a que el espacio de estas funciones incluye un término constante y un termino lineal. En la expresión (42) sólo $M + 2$ de los términos B-splines son usados; la constante es absorbida en la función de riesgo base y el término lineal es especificado separadamente para facilitar la especificación de la hipótesis de un efecto lineal, siempre que la parametrización resultante sea de rango completo.

Para un spline cúbico, $s_k$, la función de penalización frecuentemente utilizada es dada por (O'Sullivan, 1986)

$$\frac{1}{2}\lambda_k \int [s_k''(z_k)]^2 dz, \tag{44}$$

donde $\lambda_k$ es el parámetro de suavizamiento que controla el grado de suavidad, con $\lambda_k = 0$ para ninguna penalización y $\lambda_k \to \infty$ forzando $\lambda_k = \theta_{k,0} z_k$.

La novedad que introducen los P-splines es que la penalización es discreta y que se penalizan los coeficientes directamente, en vez de penalizar la curva, lo que reduce la dimensionalidad del problema en comparación al método de suavizamiento spline.

En el modelo (42) con componentes no lineales parametrizados por (43) y utilizando la penalización (44) en la función de verosimilitud parcial, el problema de modelamiento se reduce a la estimación de los parámetros y a realizar la prueba de hipótesis, para evaluar el efecto de las covariables en la función de riesgo.

### 4.5.2. Estimación de los parámetros.

Sean $\vartheta_k = (\theta_{k,0}, \theta_k)$ los parámetros de las componentes no lineales del modelo (42), con términos no lineales denotados como $\theta_k = (\theta_{k,1}, \ldots, \theta_{k,M+2})$. En la función de penalización solo aparecen los parámetros $\theta_k$ y (44) es una función cuadrática en los parámetros, por lo que se puede escribir como (Gray, 1992)

$$\frac{1}{2}\lambda_k \theta_k' P_k \theta_k = \frac{1}{2}\lambda_k \vartheta_k' P_k^* \vartheta_k,$$

donde $P$ es una matriz definida positiva que es una función solamente de la localización de los nodos y $P^*$ es una matriz $(M+3) \times (M+3)$ con ceros en la primera fila y columna y $P$ en el resto de la matriz.

Denotando al conjunto de parámetros $\beta$ y $\vartheta$ por $\eta = (\beta, \vartheta)$, y $\ell(\eta)$ el logaritmo de la función de verosimilitud parcial (Cox, 1972) para el modelo (42) con $s_k$ parametrizado por (43). El logaritmo de la verosimilitud parcial penalizado ($\ell_p(\eta)$) es de la forma,

$$\ell_p(\eta) = \ell(\eta) - \frac{1}{2}\sum_{k=1}^{q} \lambda_k \theta_k' P_k \theta_k, \tag{45}$$

donde $\theta_k$ es un vector de parámetros asociados a un spline $s_k$ y $P_k$ es una matriz definida no negativa, $\lambda_k$ es el parámetro de suavizamiento que controla el grado de suavidad de la curva a través de su segunda derivada.

Para estimar los parámetros del modelo se necesita especificar los parámetros de suavizamiento $\lambda_k$ o los grados de libertad del modelo. Otro procedimiento más sofisticado sería utilizando una selección automática basado en la validación cruzada (O'Sullivan 1988).

El estimador de máxima verosimilitud parcial penalizado de los parámetros, $\eta = (\beta, \vartheta)$, se obtienen maximizando el logaritmo de la función de verosimilitud parcial penalizado (45), como una solución de las ecuaciones generadas por las derivadas parciales del logaritmo de la verosimilitud respecto de los parámetros, $\frac{d\ell_p(\eta)}{d\eta} = 0$, utilizando métodos iterativos como Newton-Rapson.

Obtenidos los estimadores máximo verosímiles de los parámetros se puede realizar las pruebas de hipótesis sobre los parámetros, tanto de los términos lineales como no lineales del modelo aditivo, basada en la distribución asintótica de los estimadores de máxima verosimilitud.

Sea $U(\eta,\lambda) = \left(\frac{d\ell_p(\eta)}{d\eta}\right)$ la función score de la verosimilitud parcial penalizado e $I(\eta,\lambda) = -\left(\frac{d^2\ell_p(\eta)}{d\eta d\eta'}\right)$ la matriz de información. Bajo la normalidad asintótica de los estimadores de máxima verosimilitud parcial penalizado, los estadísticos de prueba pueden ser construidas exactamente como en el análisis de verosimilitud ordinario utilizando estas cantidades para realizar las distintas pruebas de hipótesis sobre los parámetros del modelo aditivo parcialmente lineal (42).

### 4.5.3. Prueba de hipótesis e inferencia.

En el modelo de la forma (42) existen tres hipótesis de prueba, aunque dos de ellas referidas a los componentes no lineales son de interés particular:

- Hipótesis sobre el efecto global, $H_0 : \eta = 0$ vs. $H_1 : \eta \neq 0$
- Hipótesis sobre el efecto de la $k$-ésima covariable, $H_{01_k} : \vartheta_k = 0$ vs. $H_{01_k} : \vartheta_k \neq 0$
- Hipótesis sobre el efecto lineal de la $k$-ésima covariable, $H_{02_k} : \theta_k = 0$ vs. $H_{02_k} : \theta_k \neq 0$

Para una muestra, $n$, suficientemente grande, sea $(\hat{\beta}, \hat{\vartheta})$ los valores que maximizan el logaritmo de la verosimilitud, $\ell_p(\eta)$. Bajo la distribución asintótica de los estimadores de máxima verosimilitud, se pueden construir diferentes estadísticos de prueba para el contraste de hipótesis sobre los parámetros del modelo (42).

**Prueba de hipótesis sobre el efecto de las componentes del modelo**:

Sea la hipótesis $H_0 : \eta = 0$. Sean $\hat{\eta}$ el estimador de máxima verosimilitud parcial del vector de parámetros, $U(\eta,\lambda)$ la función score y $I(\eta,\lambda) = (I(\eta) + \lambda P^*)$ la matriz de información, donde $I(\eta)$ es la usual matrix de información de la verosimilitud parcial. Los estadísticos de prueba para contrastar la hipótesis, basados en la distribución asintótica de los estimadores son dados por:

- Estadístico de prueba tipo Wald: $Q_W = \hat{\eta}'(I(\hat{\eta},\lambda))\hat{\eta}$
- La prueba de razón de verosimilitud penalizado: $Q_{LR} = 2[\ell(\hat{\eta}) - \ell(\eta_0)]$
- Estadístico de prueba score penalizado: $Q_{Sc} = U(\eta_0,\lambda)'\ I^{-1}(\eta_0,\lambda)U(\eta_0,\lambda)$,

las cuales, tienen una distribución Chi-cuadrado con $d.f = traza(I(\eta) * I^{-1}(\eta,\lambda))$ grados de libertad (Gray, 1992).

**Prueba de hipótesis sobre el efecto de las componentes no lineales**:

Sea la hipótesis $H_0 : \vartheta = 0$. Sea $\hat{\beta}_0$ el estimador de la máxima verosimilitud parcial para $\beta$ cuando $\vartheta = 0$. Sea $U(\beta, \vartheta) = (U'_\beta(\beta, \vartheta), U'_\vartheta(\beta, \vartheta))$ el score de la verosimilitud parcial e $I$ la matriz de información de la verosimilitud parcial no penalizado, con subíndices denotando las submatrices, tales como $I_{\vartheta\vartheta}$ para las derivadas con respecto a $\vartheta$. Note que $\left(\frac{d\ell_p(\hat{\beta}_0,0)}{d\vartheta}\right) = U_\vartheta(\hat{\beta}_0, 0)$, y que el negativo de la parte $\vartheta\vartheta$ de la matriz de la segunda derivada del logaritmo de la verosimilitud penalizado es $I_{\vartheta\vartheta} + \lambda P^*$, con el otro componente los componentes correspondientes de $I$.

Por analogía con el usual procedimiento de verosimilitud paramétrica, los tres diferentes estadísticos de prueba son dados por:

- Estadístico de prueba tipo Wald: $Q_W = \hat{\vartheta}'(I_{\vartheta\vartheta/\beta} + \lambda P^*)\hat{\vartheta}$.
- Prueba de razón de verosimilitud penalizado: $Q_{LR} = 2[\ell_p(\hat{\beta}, \hat{\vartheta}) - \ell_p(\hat{\beta}_0, 0)]$.
- Estadístico de prueba score penalizado: $Q_{Sc} = U'_\vartheta(\hat{\beta}_0, 0)(I_{\vartheta\vartheta/\beta} + \lambda P^*)^{-1} U_\vartheta(\hat{\beta}_0, 0)$, donde $I_{\vartheta\vartheta/\beta} = I_{\vartheta\vartheta} - I_{\vartheta\beta} I_{\beta\beta}^{-1} I_{\beta\vartheta}$.

Donde, las estadísticas de prueba tienen una distribución Chi-cuadrado con $d.f = traza(I_{\vartheta\vartheta/\beta} * (I_{\vartheta\vartheta/\beta} + \lambda P^*)^{-1})$ grados de libertad (Gray, 1992).

**Prueba de hipótesis sobre el efecto lineal de las componentes no lineales**:

La construcción de los estadísticos de prueba para la hipótesis de que la $k$-ésima covariable presenta un efecto lineal, $H_{0_k} : \theta = 0$, se puede hacer exactamente lo mismo, pero con $\theta_0$ incluido con $\beta$ en lugar de $\vartheta$.

Sea la hipótesis $H_0 : \theta = 0$. Sea $\hat{\beta}_0^*$ el estimador de la máxima verosimilitud parcial para $\beta^*$ ($\beta^* = (\beta, \theta_0)$) cuando $\theta = 0$.

Sea $U(\beta^*, \theta) = (U'_{\beta^*}(\beta^*, \theta), U'_\theta(\beta^*, \theta))$ el score de la verosimilitud parcial e $I$ la matriz de información de la verosimilitud parcial no penalizado, con subíndices denotando las submatrices, tales como $I_{\theta\theta}$ para las derivadas con respecto a $\theta$. Note que $\left(\frac{d\ell_p(\hat{\beta}_0^*,0)}{d\theta}\right) = U_\theta(\hat{\beta}_0^*, 0)$, y que el negativo de la parte $\theta\theta$ de la matriz de la segunda derivada del logaritmo de la verosimilitud penalizado es $I_{\theta\theta} + \lambda P^*$, con el otro componente los componentes correspondientes de $I$.

Similarmente al procedimiento anterior, los tres diferentes estadísticos de prueba son dados por:

- Estadístico de prueba tipo Wald: $Q_W = \hat{\theta}'(I_{\theta\theta/\beta^*} + \lambda P^*)\hat{\theta}$.
- Prueba de razón de verosimilitud penalizado: $Q_{LR} = 2[\ell_p(\hat{\beta^*}, \hat{\theta}) - \ell_p(\hat{\beta}_0^*, 0)]$.
- Estadístico de prueba score penalizado: $Q_{Sc} = U'_\theta(\hat{\beta}_0^*, 0)(I_{\theta\theta/\beta^*} + \lambda P^*)^{-1} U_\theta(\hat{\beta}_0^*, 0)$, donde $I_{\theta\theta/\beta^*} = I_{\theta\theta} - I_{\theta\beta^*} I^{-1}_{\beta^*\beta^*} I_{\beta^*\theta}$.

Así mismo, estas estadísticas de prueba tienen una distribución Chi-cuadrado con $d.f = traza(I_{\theta\theta/\beta^*} * (I_{\theta\theta/\beta^*} + \lambda P^*)^{-1})$ grados de libertad (Gray, 1992).

En general, el contraste de hipótesis sobre el fecto de las componentes del modelo (42)) se rechazan para valores grandes de $Q_W$, $Q_{LR}$ y $Q_{Sc}$.

**Prueba de hipótesis sobre el efecto individual de las componentes lineales del modelo**:

Por otro lado, para contrastar la hipótesis $H_0 : \beta_j = 0$ versus $H_0 : \beta_j \neq 0$, se puede utilizar la estadística de prueba tipo Wald, $(\hat{\beta}_j / \sqrt{\hat{v}(\hat{\beta}_j)})^2$, donde $v$ es la varianza del estimador del parámetro $\beta_j$. Este esdístico tiene una distribución Chi-cudrado con un grado libertad para una covariable contínua o binaria y $r - 1$ grados de libertad para una covariable con $r$ categorías o niveles.

En el programa R, la función coxph y pspline de la librería package survival, permiten realizar el ajuste del modelo de Cox con los componentes no lineales aproximado mediante splines penalizados. La estimación de los parámetros se realiza con 10 nodos basado en los cuantiles. Los estadísticos de prueba disponibles son: la prueba de razón de verosimilitud y la prueba tipo-Wald.

## 4.6. Métodos de diagnóstico en el modelo de Cox

En un modelo de regresión lineal es fácil definir un residuo. Sin embargo, en el modelo de regresión para datos de supervivencia la definición del residuo no es tan clara. Una serie de residuos se han propuesto para el modelo de Cox, que son útiles para examinar los diferentes aspectos del modelo (Klein y Moeschberger (1997), Therneau y Grambsch (2000)).

En el modelo Cox (17), las restricciones naturales suponen verificar los siguientes aspectos:

- El logaritmo de la razón de riesgo no depende del tiempo; es decir el modelo es de riesgo constante ($ln(\lambda(t,x)/\lambda_0(t)) = ln(HR) = \beta_1 X_1 + \ldots + \beta_p X_p$ ).
- El riesgo de un individuo es proporcional al riesgo de otro individuo (riesgo proporcional). Es decir, la razón de riesgo para dos individuos con los mismos valores para las covariables pero que reciben diferentes tratamientos, es simplemente $e^{\beta_l}$, asumiendo que $x_l$ es indicador de tratamiento.
- El logaritmo de la razón de riesgo y las covariables se relacionan linealmente (forma funcional lineal); es decir el efecto de las covariables se relaciona linealmente con $ln(HR)$ ($ln(HR) = \beta_1 X_1 + \ldots + \beta_p X_p$ ).

Una medida para evaluar la suposición de riesgos proporcionales puede ser realizada mediante métodos numéricos o aproximaciones gráficas.

Aquí describimos brevemente los procedimientos basados en la gráfica de los residuos de Schoenfeld escalado (Grambsch y Therneau, 1994), el estadístico de prueba de no proporcionalidad de Therneau y Grambsch (2000) y los residuos martingala.

### 4.6.1. Verificación del supuesto de riesgos proporcionales.

#### *4.6.1.1. Método gráfico basado en los residuos de Schoenfeld escalado.*

Schoenfeld (1982) propone unos residuos para verificar la suposición de riesgo proporcional. Estos residuos son conocidos como residuos de Schoenfeld. El residuo de Schoenfeld es la diferencia entre el valor observado y esperado de la covariable en momento del tiempo (Therneau y Grambsch 2000).

Sea el modelo de Cox extendido con efectos que varían en el tiempo (Extended Cox model with time-varying coefficients),

$$\lambda(t) = Y(t)\lambda_0(t)\exp(X'(t)\beta(t)) \tag{46}$$

donde $\beta(t)$ es el efecto tiempo dependiente, que cuando no es constante, el impacto de una o más covariables en el riesgo puede variar sobre el tiempo. Pero la restricción $\beta(t) = \beta$ implica riesgo proporcional, por tanto, la gráfica de $\beta(t)$ versus tiempo sería una línea constante.

Los residuos de Schoenfeld permiten detectar la variación en el tiempo para un predictor de interés. En ausencia de empates estos residuos son iguales a la diferencia entre el vector de covariables observados y esperados para un evento en el tiempo $t_k$ $(k = 1, \ldots, d)$,

$$\widetilde{r}_k = \widetilde{X}_k - E(\widetilde{X}_k/R_k),$$

donde $E(\widetilde{X}_k/R_k) = \left(\frac{\sum_{l \epsilon R_k} \widetilde{X}_l \exp(\widetilde{\beta} X_l)}{\sum_{l \epsilon R_k} \exp(\widetilde{\beta} X_l}\right)$.

En la presencia de $p$ covariables, los residuos de Schoenfeld $\widetilde{r}$ forman una matriz $d \times p$, donde cada covariable $p$ tiene un coeficiente estimado para cada evento del tiempo, $\beta_{kp}$.

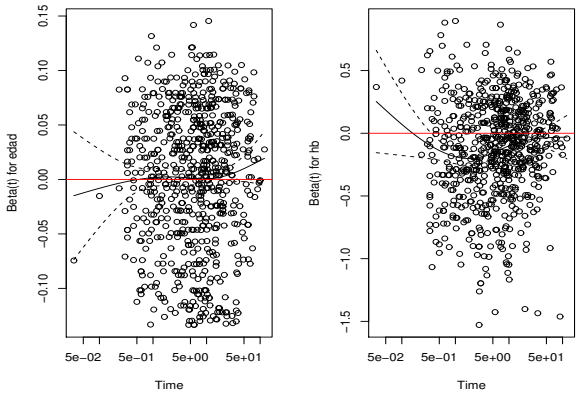
Los residuos de Schoenfeld escalados $(rs)$ se definen como el producto de la inversa del estimador de la matriz de varianza - covarianza del $k$-ésimo residuo de Schoenfeld y el $k$-ésimo residuo de Schoenfeld. Grambsch y Therneau (2000) muestran que $E(rs_{kp}) + \hat{\beta}_p \approx \beta_p(t_k)$, donde $rs_k$ son los residuos de Schoenfeld escalados y $\hat{\beta}$ es un coeficiente estimado del modelo de Cox.

Esto sugiere graficar $rs_k + \hat{\beta}_p$ versus tiempo, o alguna función de tiempo $g(t)$, como un método para visualizar la forma funcional de la variación en el tiempo de los residuos Schoenfeld escalados para una covariable específica. En el supuesto de riesgos proporcionales, los residuos se distribuyen alrededor de la línea horizontal para un coeficiente $\beta_p$ constante.

Para facilitar la interpretación de estos gráficos se superpone una curva de ajuste, utilizando alguna función de ajuste local como lowess o loess (Cleveland, 1981). Si se cumple la hipótesis de riesgo proporcional, los residuos deberían agruparse de forma aleatoria a ambos lados del valor 0 del eje Y, y la curva ajustada debería ser próxima a una línea recta.

En la gráfica 6, se muestra la relación entre $rs_k + \hat{\beta}_p$ y $g(t) = log(t)$, para los datos de la edad y la hemoglobina. Las líneas negras corresponden a la curva de ajuste de los residuos Schoenfeld escalado $\pm$ error estándar aproximado mediante lowess y la línea roja es la recta horizontal en el punto 0 del eje de las ordenadas.

En esta gráfica se observa que el efecto de la edad no varía en el tiempo, ya que los residuos se aproximan a la línea horizontal en el punto 0 del eje Y, lo que significa el cumplimiento de riesgo constante. En cambio el efecto de la hemoglobina disminuye y después se incrementa a lo largo del tiempo, lo que contradice la suposición de riesgo constante a lo largo del tiempo de un modelo de Cox correctamente especificado.




GRÁFICA 6. Residuos Schoenfeld escalado vs. $g(t) = log(t)$, para la edad y Hb.

*4.6.1.2. Test de no-proporcionalidad de Therneau y Grambsch.*

Siguiendo la aproximación gráfica, Grambsh y Therneau (1994) introducen una versión del test del score basado en los residuos de Schoenfeld escalados.

Escrito $\beta(t)$ como una función de regresión en $g(t)$, los coeficientes tiempo dependientes del modelo de Cox extendido (46) pueden ser escritos como,

$$\beta_j(t) = \beta_j + \theta_j(g_j - \bar{g}_j), j = 1, \ldots, p \tag{47}$$

donde $\bar{g}_j$ es la media de la $g_j$ (función de tiempo especificado). Una típica aplicación de este tipo de prueba es para $g_j = log(t)$.

En la expresión (47) el interés es realizar una prueba de hipótesis sobre la hipótesis de riesgo proporcional global $H_0 : \theta = 0$ y para una covariable específica $H_{0j} : \theta_j = 0$, $j = 1, \ldots, p$.

Para realizar estas pruebas de hipótesis, Therneau y Grambsch (2000) introducen dos estadísticos de prueba, una para la global y otro para una covariable especifica.

- Para la hipótesis global, $H_0 : \theta = 0$, el estadístico de prueba de riesgo proporcional para todas las $p$ covariables es

$$T = \frac{(g - \bar{g})' S^* I S^{*'} (g - \bar{g})}{d \sum (g_k - \bar{g})^2},$$

donde $S^*$ es la matriz de los residuos de Schoenfeld escalados, $I$ es la matriz de información y $d$ es eventos de tiempo.

- Para la hipótesis de una covariable específica $H_{0j} : \theta_j = 0$, el estadístico de prueba de riesgo proporcional es

$$T_j = \frac{(\sum(g_k - \bar{g})rs_{kj})^2}{dI_{jj}\sum(g_k - \bar{g})^2},$$

donde $I_{jj}$ es el elemento de la matriz de información para la $j$-ésima covariable y $d$ son los eventos de tiempo. Este estadístico se distribuye asintóticamente como una $\chi^2_1$.

Por otro lado, sean los coeficiente de regresión tiempo-dependiente del modelo de Cox extendido (46) que puede ser escrito como

$$\beta_p(t) = \beta_p + \theta_p g_p(t), \tag{48}$$

donde $g_p(t)$ es una función de tiempo especificado previamente.

Cuando las $g's$ son funciones conocidas, entonces el modelo con coeficientes (48) es aún el modelo de Cox. Por tanto, el estimador de los parámetros se puede obtener mediante la función de verosimilitud partial y realizarse las pruebas de hipótesis sobre las componentes tiempo-dependientes utilizando el test de score, la prueba de razón de verosimilitud o la prueba de Wald. (Martinussen y Scheike (2006), Therneau y Grambsch (2000)).

Sea $U = (U_1', U_2')$ la función de puntaje, donde la primera componente es la derivada de la verosimilitud parcial con respecto a $\beta$ y la segunda componente respecto a $\theta$. Sea $I_{kl}$ $(k, l = 1, 2)$ la matriz de información empírica definida como una matriz de bloques reflejando dos vectores de parámetros, .

Para realizar la prueba de hipótesis sobre $H_{02} : \theta = 0$, $\theta = (\theta_1, \ldots, \theta_p)$ de los parámetros de la expresión (48), se puede utilizar el estadístico de prueba global (score test) para los efectos tiempo-dependientes definida como

$$T(G) = U_2'(\hat{\beta}_p, 0)I_{22}^{-1}(\hat{\beta}, 0)U_2(\hat{\beta}_p, 0),$$

donde $\hat{\beta}$ denota el estimador de máxima verosimilitud parcial. Este estadístico se distribuye como una $\chi^2$ con $p$ grados de libertad bajo la hipótesis nula.

Para los datos de la gráfica de los residuos de Schoenfeld escalados vs. logaritmo del tiempo (Gráfica 6), los resultados (valores $p$) del estadístico de prueba de no proporcionalidad de Therneau y Grambsch indican que el efecto de la edad en la función de riesgo es constante (p = 0.359) y en cambio el efecto de la hemoglobina es de riesgo no constante (p = 0.009).

En el package R, la función *cox.zph* permite realizar la gráfica de los residuos de Schoenfeld escalados versus una transformación de la función tiempo ($g(t)$) y obtener los estadísticos de prueba individual y global. Las transformación de la función de tiempo disponibles son la identidad $g(t) = t$, $g(t) = log(t)$, rangos de eventos de tiempo y por defecto es 1-KM (KM: es el estimador de Kaplan-Meier).

### 4.6.2. Verificación de la forma funcional lineal.

En el modelo de Cox, uno de los residuos muy usuales para verificar la forma funcional del efecto de las covariables en la función de riesgo son los residuos basados en martingalas. Barlow y Prentice (1988) provee el marco básico de los residuos martingala y el posterior trabajo de Therneau, Grambsch y Fleming (1990).

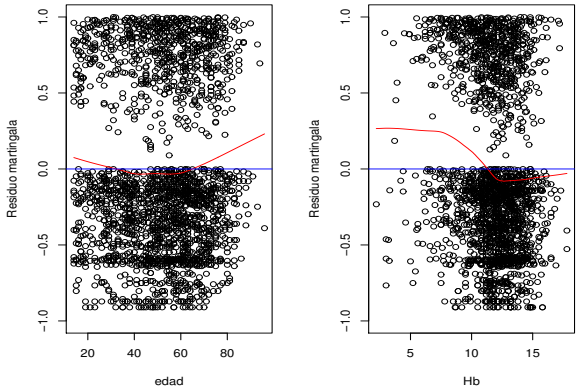
Sea $N_i(t)$ el proceso de conteo (número de eventos observados) y la función de intensidad acumulada $\Lambda_i(t) = \int_0^t \lambda_s ds$, con información adicional en términos de $p$ covariables $X_i$. Los residuos martingala $M_i(t)$ se definen como la diferencia entre los procesos de conteo observados y esperados, $M_i = N_i(t) - E_i(t)$, donde $E_i(t) = \Lambda_i(t)$.

Bajo la función de intensidad de la forma (25) y usando los estimadores del modelo de Cox se pueden estimar los residuos martingala, $M_i(t)$. Por otro lado, los residuos martingala se pueden construir basándose a su vez en los denominados residuos de Cox-Snell ($rc_i$), $\widehat{rm}_i = \delta_i - \widehat{rc}_i$, donde $\delta_i$ es 1 si ocurre el evento, 0 caso contrario y los residuos de Cox-Snell se definen como el estimador de la función de intensidad acumulado, $\widehat{rc}_i = \hat{\Lambda}(t_i)$, $i = 1, \ldots, n$.

Si la muestra es grande, la suma de los residuos martingala es cero, son no correlacionados y el valor esperado es cero. Sin embargo, no se distribuyen de forma simétrica en torno a cero, aunque el modelo sea correcto, lo que dificulta la interpretación de los gráficos. La gráfica de los residuos martingala versus la covariable, bajo el supuesto de efecto lineal en el modelo, deben verificar que los residuos se distribuyen alrededor de un punto del eje $y$, sin que sugiera una curva de ajuste de forma funcional no lineal.

En el gráfica 7 se muestra los residuos martingala versus la edad y hemoglobina. Las líneas negras corresponden a la curva de ajuste de los residuos aproximado mediante lowess y la línea roja es la recta horizontal en el punto 0 del eje $y$. En esta gráfica se observa que el efecto de la edad y de la hemoglobina no es lineal; ya que la curva de ajuste de los residuos versus la edad y hemoglobina (líneas rojas) no se aproximan a una línea horizontal. Los cuales, significan que el efecto de las covariables en la función de riesgo presentan una forma funcional no lineal.

Por otro lado, Lin et al. (1993) y Wei (1984) sugieren una importante clase de estadísticos de prueba basado en la suma acumulada de los residuos martingala. Estos estadísticos son diseñados para investigar las diferentes salidas del modelo, incluyendo errores de especificación de la función de enlace y la forma funcional de las covariables (Martinussen y Scheike, 2006).




GRÁFICA 7. Residuos martingala vs. edad y Hb.

Las martingalas bajo el supuesto de riesgo proporcional del modelo de Cox se pueden escribir como

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(X_i'\beta) d\Lambda_0(s).$$

En el cual, usando los estimadores del modelo de Cox se puede estimar $M_i(t)$ como

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(X_i'\hat{\beta}) d\hat{\Lambda}_0(s).$$

La idea ahora es mirar las diferentes funcionales de estos residuos estimados y ver si se comportan como debería bajo el modelo propuesto.

Lin et al. (1993) define un proceso de residuales acumulativo bi-dimensional como

$$M_c(t, z) = \int_0^t K_z^t(s) d\hat{M}(s),$$

donde $K_z(t)$ es una matriz $n \times 1$ con elementos $I(X_{i1} \leq z)$ para $i = 1, \ldots, n$, centrándose aquí en la primera covariable continua $X_1$. En este caso, los residuos martingala son agrupados de forma acumulativa respecto al tiempo de seguimiento y valores de la covariable. Para resumir éstas se puede integrar sobre el periodo de tiempo y conseguir un proceso únicamente en $z$, $M_c(z) = \int_0^t K_z'(t) d\hat{M}(t)$, el cual puede ser graficado contra $z$.

Para evaluar el proceso observado como inusual bajo el modelo propuesto, se puede graficar este a lo largo del tiempo como una realización bajo el modelo. Para mejorar aún más la objetividad de la técnica gráfica, se puede completar con con un estadístico de prueba llamada test de supremo de $M_c(t)$; el cual, mide el extremo del proceso observado.

Un valor demasiado grande de este test sugiere que la forma funcional lineal del efecto de la covariable es inapropiada, lo que significa que las variables no podrían entrar en el modelo en la escala original; por tanto, esta variable requiere algún tipo de trasformación para ser incluido en el modelo.

En el package R, una función importante asociada a los objetos del tipo *coxph* es la función *residuals*, o en su formato más corto *resid*. Esta función permite calcular los residuos martingala y verificar la forma funcional del efecto de las covariables basado en la gráfica de los residuos verus las covariables, que bajo el supuesto de efecto lineal los residuos deben distribuirse formando una especie "nube de puntos" sin que sugiera una curva que indique falta de ajuste.

Otros residuos disponible bajo la función *rsid* son: los residuos de puntaje o score (utilizado para verificar la influencia individual y para la estimación robusta de la varianza), de desvío o deviance (utilizado para la detección de valores atípicos (outliers)) y de Schoenfeld (utilizado para verificar el supuesto de riesgo proporcional), no presentados aquí.

Para el segundo criterio, en el package R, la función cox.aalen permite realizar la simulación de los residuos y obtener la gráfica de los residuos acumulados vs. las covariables continuas, así como el test supremo, no presentados aquí.

# Capítulo 5
# Factores pronósticos en LNH

En este capítulo se presenta los resultados de la aplicación de los modelos descrito en la sección 4.4 y 4.5, para determinar el efecto y la forma funcional no lineal del efecto (forma funcional) de las covariables (factores pronósticos) para la supervivencia global en un grupo de pacientes con LNH diagnósticados y tratados en el INEN entre 1990 a 2002, utilizarándo regresión splines y suavizamiento splines penalizado (P-splines) con bases B-splines cúbico para aproximar la forma funcional del efecto de las covariables en el modelo de Cox. Previamente se presenta la descripción de los datos, luego se presenta los resultados de la aplicación del modelo de Cox clásico, modelo de Cox con regresión splines y modelo de Cox con P-splines.

## 5.1. Descripción de los datos

En este trabajo se analizan los datos de 2160 pacientes mayores o iguales a 14 años de edad con diagnóstico de linfoma no Hodgkin (LNH), que fueron diagnósticados y tratados en el Instituto Nacional de Enfermedades Neoplásicas (INEN), Lima-Perú, entre 1990 y 2002. Así mismo, cabe resaltar que los datos corresponden a una sub-base del estudio retrospectivo clínico, patológico y epidemiológico del LNH, donde uno de los objetivos del estudio fue determinar los factores pronósticos para la supervivencia global de esta patología.

### 5.1.1. Descripción de las variables.

Durante el estudio retrospectivo, se recopilaron los datos relacionados a los aspectos epidemiológicos, clínicos y patológicos al momento del diagnóstico, las características del tratamiento y seguimiento. De los cuales, para los propósitos de nuestro aplicación y análisis sólo utilizamos los datos relacionados a las características del paciente y del tumor (características clínicas), así como del estado de seguimiento (tiempo de supervivencia y condición actual de seguimiento).

Las seguientes características clínicas (covariables) fueron incluidos en el análisis de los factores pronósticos: la edad, género, estado de performante, localización de la enfermedad, estadio clínico (EC), síntomas B, hemoglobina (Hb), leucocitos, linfocitos y la deshidrogenasa láctica (DHL) sérica. No se incluye en el análisis el tamaño del tumor, número de ganglios afectados, sitios extragaglionares y el sitio de metástasis debido a que estas variables ya están reflejadas en el estadio clínico. Así mismo, el tipo de linfoma y el genotipo debido a que el diagnóstico fue realizado con tres criterios de clasificación histopatológica diferentes que corresponden a tres periodos (Rappaport y Kiel, formulación de trabajo (WF) y la clasificación REAL) y la $\beta 2M$ debido a que fue determinado en pocos pacientes.

En la Tabla 1, se describe las variables incluidas en el análisis de los factores pronósticos para la supervivencia global de los pacientes con LNH.

Tabla 1. Descripción de las variables de estudio.

| Variables | Descripción | Categorías de la variable |
|---|---|---|
| Edad | Edad al diagóstico en años | variable contínua |
| Género | Sexo del paciente | femeno o masculino |
| Zubrod | Estado general según la escala ECOG | 0,1,2,3 o 4 |
| Primario | Localización de la enfermedad | ganglionar o extraganglionar |
| Estadio | Estadio clínico de la enfermedad (según la clasificación Ann Arbor) | I, II, III o IV |
| Síntomas | Presencia de síntomas | A o B |
| Hemoglobina | Nivel de hemoglobina en $g/dl$ | variable continua |
| Leucocitos | Numero de leucocitos por $mm^3$ | variable discreta |
| Linfocitos | Porcentaje de linfocitos | variable discreta |
| DHL | Nivel de DHL en UI/L | Variable discreta |

B: fiebre, sudoración nocturna o baja de peso sin causa alguna. A: no sintomas B.

La edad avanzada se relacionan con un peor pronóstico, ya que va asociada a una mayor morbimortalidad; según la mayoría de los grandes estudios, los pacientes mayores de 60 0 70 años de edad pueden estar relacionado a un peor pronóstico. La variable zubrod se refiere al compromiso que la enfermedad le produce en el estado general del paciente, conocido también como estado de performance. Según la escala ECOG (East Cooperative Oncology Group) el valor varia de 0 a 4; el valor 0 indica sin deterioro (paciente con actividad normal y asintomático) y el valor 4 deterioro completo (postrado permanentemente o terminal).

El estadío clínico (EC) hace referencia a la extensión de la enfermedad; EC I y II significa tumor localizado (enfermedad temprana), mientras EC III y IV indica enfermedad localmente avanzada a enfermedad extendida a otros órganos (enfermedad avanzada).

La hemoglobina es un indicadores del estado anémico; según el criterio clínico, un valor menor que lo normal indica riesgo de desnutrición. Los leucocitos y los linfocitos pueden estar relacionados con el proceso inflamatorio, como una reacción al proceso de crecimiento anormal de las celulas (crecimiento tumoral); según el criterio clínico, un valor mayor que los valores normales pueden estar relacionados al proceso de crecimiento tumoral. La DHL sérica es un marcador relacionado con el proceso de crecimiento tumoral; según el criterio clínico, los niveles superiores a los normales puede estár relacionado con la actividad tumoral.

En la práctica clínica, los valores de los parámetros de laboratorio (exámenes de laboratorio) se clasifican, según los criterios clínicos, en valores normales o anormales (bajo o elevado). Por otro lado, también es una de las prácticas habituales, la categorización de algunas variables continuas y se utiliza estas variables categorizadas para analizar su relación con la supervivencia.

En la mayoría de las series publicadas, con datos categorizados, los pacientes con hemoglobina baja ($<12g/dl$), número de leucocitos elevados ($>10$ mil por $mm^3$), porcentaje de linfocitos elevados ($>40\,\%$) y nivel de DHL elevada ($>240UI/L$) se consideran, particularmente, como factores pronósticos para una pobre supervivencia de los pacientes con LNH.

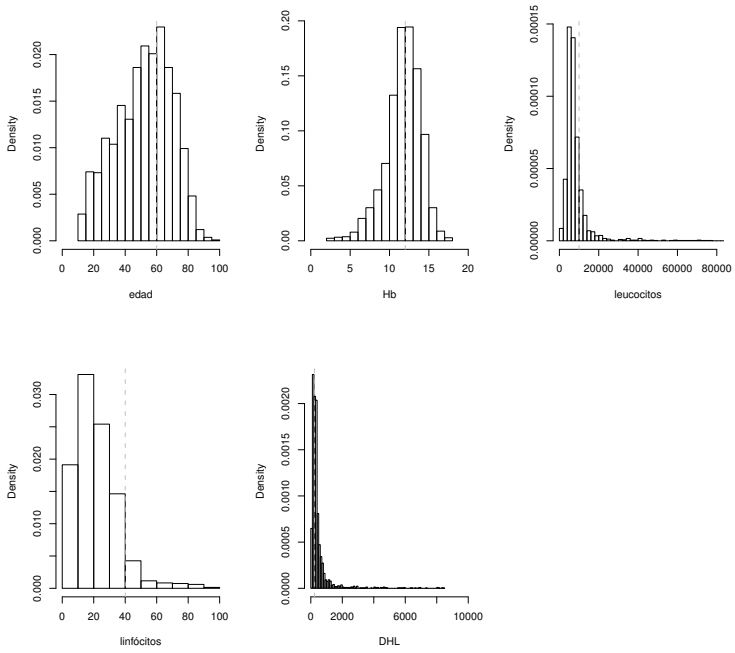
Sin embargo, si los supuestos del modelo con datos categorizados no se cumplen es preferible analizar los datos en su escala original. En la subsección seguiente se describe las covariables continuas en su escala original.

### 5.1.2. Distribución de las variables contínuas.

En la Gráfica 8 se muestra la distribución de las covariables continuas. Las gráficas muestran que todas ellas presentan asimetría; la asimetría es más pronunciada para los leucocitos y la deshidrogenasa láctica (DHL), las cuales, son posibles debido a que los valores de los leucocitos varían entre 100 a 100mil por $mm^3$ y los niveles de DHL entre 50 y $10mil\ UI/L$. Estas variables se incluyen en el modelo aplicando logarítmo natural ($ln$).

## 5.2. Características de los pacientes

En esta sección se describe de manera breve las características clínicas de los pacientes con LNH, que nos dan una idea de como se distribuyen los casos según las categorias de las variables, así mismo, se muestra la curva de supervivencia (método de Kaplan-Meier) y la curva de riesgo acumulado (método de Nelson-Aalen) de los pacientes con LNH.




GRÁFICA 8. Distribución de las variables continuas de los datos de pacientes con LNH.

### 5.2.1. Características clínicas.

En la Tabla 2 se muestra las características clínicas de los pacientes con LNH, así como el número de casos por cada una de las categorías de las variables evaluables. La edad de los pacientes varía de 14 a 96 años, alcanzando una mediana de 54 años. El 36.9 % de los pacientes eran mayores de 60 años de edad y 50.9 % fueron de sexo masculino, 27.0 % presentaban zubrod entre 2-4, 66.7 % tenían enfermedad ganglionar, 49.2 % presentaban enfermedad en estadio clínico avanzado (EC III-IV), 38.1 % habían presentado síntomas B, 48.1 % habían presentado hemoglobina baja (Hb $<12g/dl$), 17.7 % leucocitos elevados (leucocitos >10mil por $mm^3$), 7.7 % linfocitos elevados (linfocitos $> 40$ %) y 60.0 % niveles de DHL elevados ($>240\ UI/L$).

El tratamiento que habían recibido los pacientes, según la práctica clínica habitual (protocolo de tratamiento institucional), fue generalmente quimioterapia en la mayoría de los casos (91.2 %) y los restantes (8.8 %) habían recibido radioterapia y/o cirugía. El esquema de quimioterapia fue generalmente (81.6 %) CHOP (ciclofosfamida, doxorubicina, vincristina y prednisona) y los restantes otros esquemas de quimioterapia.

TABLA 2. Características clínicas de los pacientes con LNH.

| Variables | Mediana/rango | Casos | Porcentaje (%) |
|---|---|---|---|
| Edad (años) | 54.0/(14 - 96) | | |
| ≤60 | | 1363 | 63.1 |
| >60 | | 797 | 36.9 |
| Género | | | |
| Femenino | | 1061 | 49.1 |
| Masculino | | 1099 | 50.9 |
| Zubrod | | | |
| 0-1 | | 1577 | 73.0 |
| 2-4 | | 583 | 27.0 |
| Primario | | | |
| Extraganglionar | | 719 | 33.3 |
| Ganglionar | | 1441 | 66.7 |
| Estadio clínico | | | |
| I-II | | 1097 | 50.8 |
| III-IV | | 1063 | 49.2 |
| Síntomas | | | |
| A | | 1336 | 61.9 |
| B | | 824 | 38.1 |
| Hemoglobina (g/dl) | 11.8/(2.2-17.8) | | |
| ≥12g/dl | | 1122 | 51.9 |
| <12g/dl | | 1038 | 48.1 |
| Leucocitos ($10^3/mm^3$) | 6.7/(0.560-163) | | |
| ≤ 10 | | 1777 | 82.3 |
| > 10 | | 383 | 17.7 |
| Linfocitos (%) | 20/(1 - 94) | | |
| ≤40 | | 1993 | 82.3 |
| >40 | | 167 | 7.7 |
| Deshidrogenasa láctica (UI/L) | 298/(24-8440) | | |
| ≤240 | | 863 | 40.0 |
| >240 | | 1297 | 60.0 |

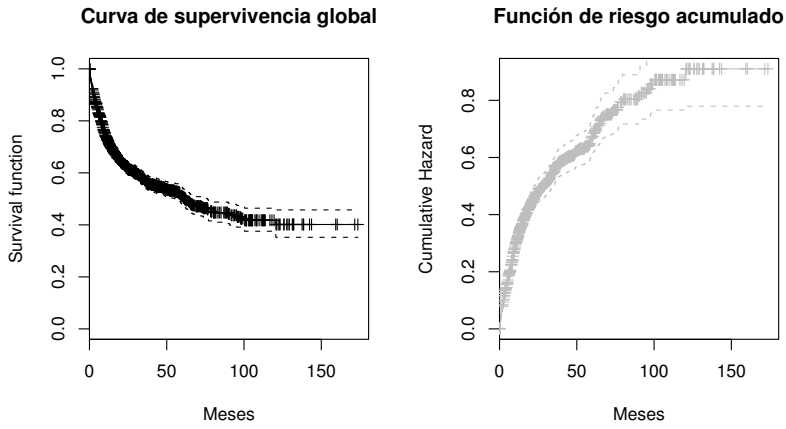
### 5.2.2. Seguimiento y supervivencia.

El tiempo de supervivencia (en meses) que es la variable a modelar en términos de las covariables, fue calculado desde la fecha de diagnóstico hasta la fecha de muerte o fecha del último control que fue registrada en la historia clínica a la fecha de la revisón de los mismos. Los pacientes fallecidos se consideraron como eventos (no censurados) y los restantes como

censurados. Por lo tanto, la variable respuesta objeto de análisis será supervivencia global de los pacientes con LNH.

De los 2160 pacientes con LNH, 709 (32.8 %) pacientes habían fallecido y la mediana de seguimiento de los pacientes restantes fue de 12.6 meses. La mediana de supervivencia global fue de 61.8 meses (IC95: 49.9 - 73.7) y la tasa de supervivencia global a 5 y 10 años de 51.2 % y 41.7 % respectivamente.

En la Gráfica 9 se muestra la curva de supervivencia global estimada mediante el método de Kaplan-Meier y la función de riesgo acumulado.




Gráfica 9. Curva de supervivencia global y riesgo acumulado de los pacientes con LNH.

De acuerdo a los resultados observados en la gráfica, se puede afirmar que la supervivencia global de los pacientes con LNH diagnósticados y tratados en el INEN es superior a 40 %; aunque la tasa a 5 y 10 años son similares a los reportados para esta patología.

En las secciones siguientes se analiza el efecto de las covariables (factores pronósticos) y su forma funcional en la supervivencia global de los pacientes con LNH mediante el modelo de Cox con regresión splines y modelo de Cox con suavizamiento splines penalizado (P-splines), utilizando bases B-splines que fueron descritos en la sección (4.4 y 4.5).

## 5.3. Aplicando el modelo de Cox clásico

En la Tabla 3 se muestran los resultados de la aplicación del modelo de Cox clásico, incluyéndose las siguientes variables: edad, género (femenino, masculino), zubrod (0-1, 2-4), primario (ganglionar, extraganglionar), estadio clínico (I-II, III-IV), síntomas (A, B), Hb, logaritmo del conteo de leucocitos (ln(leucocitos)), linfocitos y logaritmo de la DHL sérica (ln(DHL)). Los resultados incluyen los parámetros estimados ($\hat{\beta}$), la estadística de prueba y la razón de riesgo o hazard ratio (HR).

Tabla 3. Resultados bajo el modelo de Cox clásico.

| Variables | $\hat{\beta}$ | EE($\hat{\beta}$) | Z | p | HR (IC95 %) |
|---|---|---|---|---|---|
| Edad (años) | 0.005 | 0.002 | 2.210 | 0.027 | 1.01 (1.00, 1.01) |
| Género masculino | 0.202 | 0.078 | 2.590 | 0.010 | 1.22 (1.05, 1.43) |
| Zubrod 2-4 | 0.689 | 0.085 | 8.14 | <0.001 | 1.99 (1.69, 2.35) |
| Primario ganglionar | -0.046 | 0.085 | -0.550 | 0.580 | 0.96 (0.81, 1.13) |
| Estadio clínico III-IV | 0.440 | 0.086 | 5.140 | <0.001 | 1.55 (1.31, 1.84) |
| Síntomas B | 0.164 | 0.082 | 2.00 | 0.045 | 1.18 (1.00, 1.38) |
| Hemoglobina | -0.051 | 0.016 | -3.16 | 0.002 | 0.95 (0.92, 0.98) |
| ln(leucocitos) | 0.211 | 0.069 | 3.03 | 0.002 | 1.24 (1.08, 1.42) |
| Linfocitos | -0.014 | 0.003 | -4.32 | <0.001 | 0.99 (0.98, 0.99) |
| ln(DHL) | 0.255 | 0.044 | 5.73 | <0.001 | 1.29 (1.18, 1.41) |
| LR test (df) | 351.00 (10) | | | <0.001 | |
| Wald test (df) | 379.10 (10) | | | <0.001 | |
| AIC | 9597.33 | | | | |

Nota: Las categorías que no aparecen corresponden a las categorías de referencia.
Beta: Coeficiente de regresión. EE: error estándar del coeficiente de regresión.
HR: Hazard ratio.

En los resultados (Tabla 3) se observa que las covariables incluidas en el modelo presentan un efecto siginificativo en la supervivencia global de los pacientes con LNH (test de LR y Wald: p<0.001, se rechaza la hipótesis global $H_0 : \beta_1 = \ldots = \beta_p = 0$). Las covariables con efecto significativo ($p < 0.05$) fueron todas, a exepción del primario ($p = 0.580$). El efecto no significativo del primario (localización del tumor) podría estar relacionado a la heterogeneidad que involucra la localización ganglionar y extraganglionar.

Para las covariables categóricas, el efecto de estas variables implica que los pacientes de sexo masculino presentan un riesgo de mortalidad de HR=1.2 (IC5 %: 1.1-1.4) veces más

que los pacientes de sexo femenino. Los pacientes con zubrod 2-4 presentan un riesgo de mortalidad de HR=2.0 (IC95 %: 1.7-2.4) veces más que los pacientes con zubrod 0-1. Los pacientes con enfermedad avanzada (EC III-IV) presentan un riesgo de mortalidad de HR=1.6 (IC95 %: 1.3-1.8) veces más que los pacientes con enfermedad temprana (EC I-II); los cuales, condicionan el pronóstico de los pacientes.

Así mismo, para las covariables continuas, el riesgo de mortalidad se incrementa en HR=1.24 por cada unidad que incrementa el logaritmo de número de leucocitos, así mismo, el riesgo de mortalidad se incrementa en HR=1.29 por cada unidad que incrementa el logaritmo de la DHL sérica.

De aquí podemos afirmar (preliminarmente) que los factores pronósticos para la supervivencia global de los pacientes con LNH son todas las covariables, a excepción del primario; sin embargo, aún estaría pendiente la verificación de los supuestos del modelo de Cox (supuesto de riesgos proporcionales y relación lineal).

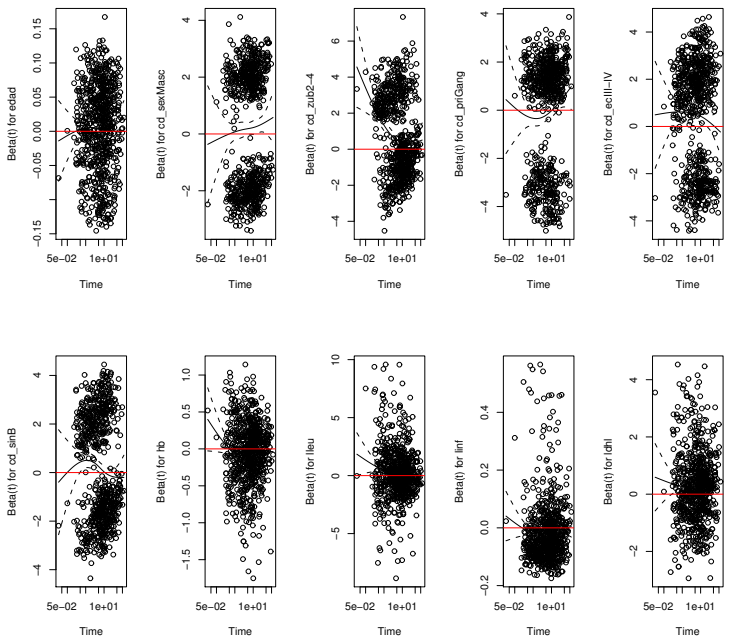
De acuerdo a los procedimientos definidos en la sección 4.6, se verifica el supuesto de riesgo proporcional basado en los residuos de Schoenfeld escalado y test de no proporcionalidad de Therneau y Grambsch y la forma funcional del efecto de las covariables en la función de riesgo basado en los residuos martingalas y la gráfica de la relación lineal entre logaritmo de la hazard ratio (log(HR)) y las covariables.

Supuesto de riesgos proporcionales (o riesgos constantes):

En la Gráfica 10 se muestra la relación entre los residuos de Schoenfeld escalados versus tiempo (ln(meses)), para cada covariable. Las líneas negras corresponde a las curva de ajuste de los residuos ($\pm$ error estándar) aproximado mediante lowess y la línea roja es la recta horizontal en el punto 0 del eje de las ordenadas. Según el método gráfico, los residuos deben aproximarse a una recta horizontal a lo largo del tiempo si se verifica el supuesto de riesgos proporcionales.

En las gráficas se observan que los residuos (línea negra) no se aproximan a una línea recta constante (líneas rojas) a lo largo del tiempo para las siguientes variables: zubrod, primario, estadio clínico, síntomas y leucocitos. Estos resultados indican que el supuesto de riesgos proporcionales (riesgos constantes) no se verifican para estas variables; los cuales, serán verificados formalmente mediante test de no-proporcionalidad.

En la Tabla 4 se muestra los resultados del test de no proporcionalidad de Therneau y Grambsch. En los resultados se observa que el efecto de las covariables no verifican el supuesto de riesgos proporcionales (riesgos constantes) para zubrod, primario, síntomas y leucocitos ($p < 0.05$), tal como fue observado gráficamente.




GRÁFICA 10. Residuos de Schoenfeld escalados vs. tiempo (ln(meses)) para modelo de Cox clásico. De izquierda a derecha, en la primera fila se muestra para la edad, género, zubrod, primario, estadio clíncio, y en la segunda fila se muestra para síntomas, Hb, log(leucocitos), linfocitos y log(DHL)

TABLA 4. Test de no proporcionalidad de Therneau y Grambsch, basado en los residuos de Schoenfeld escalados del modelo de Cox clásico.

| Variables | correlación de Pearson (rho) | $\chi^2_{1gl}$ | valor-p |
|---|---|---|---|
| Edad (años) | 0.02496 | 0.51775 | 4.72e-01 |
| Género masculino | 0.05661 | 2.37351 | 1.23e-01 |
| Zubrod 2-4 | -0.20317 | 29.56356 | 5.41e-08 |
| Primario ganglionar | 0.10805 | 8.85480 | 2.92e-03 |
| Estadio clínico III-IV | -0.06772 | 3.46333 | 6.27e-02 |
| Síntomas B | -0.10094 | 7.47280 | 6.26e-03 |
| Hb | -0.01232 | 0.10548 | 7.45e-01 |
| ln(leucocitos) | -0.07969 | 6.57062 | 1.04e-02 |
| Linfocitos | 0.02630 | 0.83625 | 3.60e-01 |
| ln(DHL) | 0.00256 | 0.00501 | 9.44e-01 |
| Global | - | 94.08964 | 7.77e-16 |

Supuesto de relación lineal:

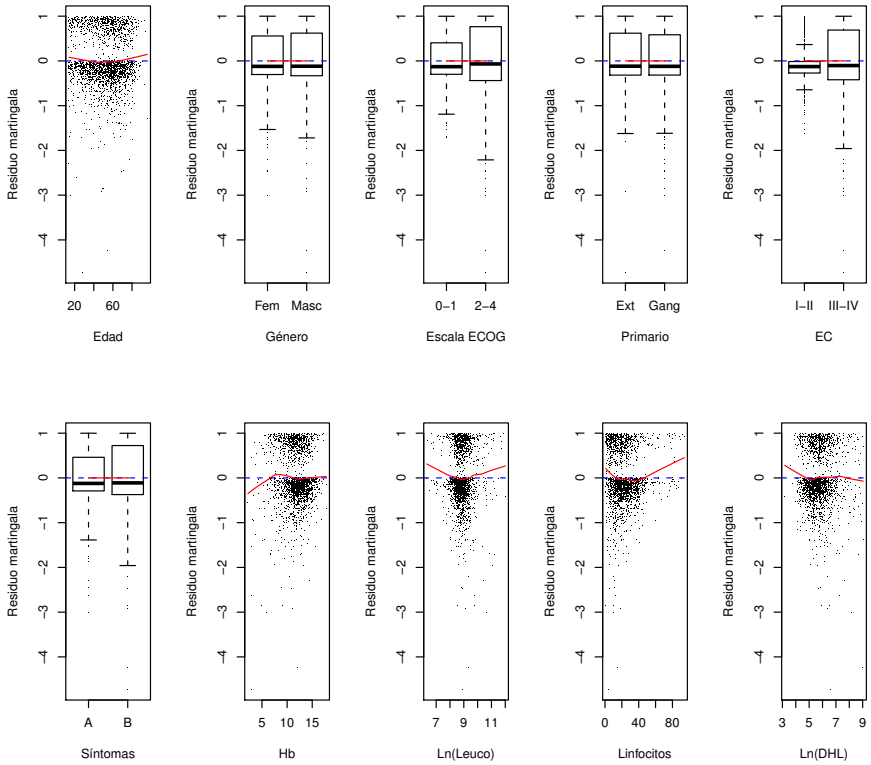
En la Gráfica 11 se muestran los residuos martingala versus las covariables. En los resultados se observan que los residuos no son constantes en relación a cada covariable contínua. Las aproximaciones de los residuos, mediante lowess (línea roja), muestran una tendencia no lineal para cada covariable contínua; los cuales indican que el efecto de las covariables no influye linealmente en la función de riesgo, y por tanto hace evidencia de una falta de ajuste de los datos.

En la Gráfica 12 se muestra la forma funcional del efecto de las covariables en el logaritmo de la razón de riesgo (ln(HR)). En los gráficos se observa que el efecto de las variables continuas en el ln(HR) no presenta una relación lineal. Estos resultados sugieren que estas variables no pueden entrar en el modelo en su escala original y por tanto necesitan de algún tipo de transformación.

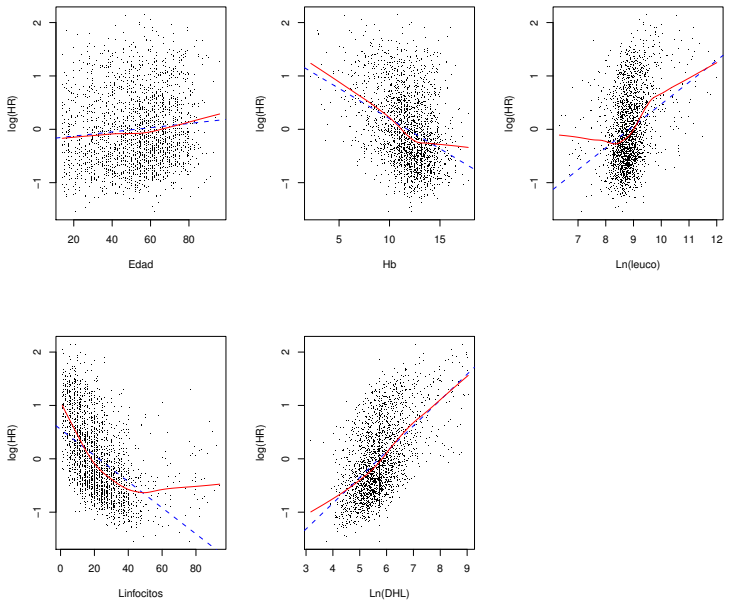
Conclusión preliminar:

De acuerdo a los resultados del test de no proporcionalidad y residuos martingala, se concluye que el supuesto de riesgos proporcionales (riesgos contantes) y el supuesto de linealidad entre el logaritmo de la razon de riesgo y las covariables no se cumplen para los datos ajustados mediante el modelo de Cox clásico; lo cual, indica que los resultados de las estimaciones de los parámetros bajo el modelo subyacente no son adecuados para los datos de los pacientes con LNH.

Si bien las trasformaciones de las covariables contínuas puede ser un recurso alternativo para linealizar el efecto de las covariables, sin embargo estos procedimiento están limitados a curvas monótonamente crecientes o decrecientes. En los análisis siguientes aproximamos la forma funcional del efecto de las covariables mediante métodos más flexibles como regresión splines y suavizamiento splines penalizado (P-splines), que en principio aproximan bién las funciones conocidas, así como la forma funcional del efecto de las covariables contínuas en los modelos aditivos.




GRÁFICA 11. Residuos martingala del modelo de Cox clásico.




GRÁFICA 12. Forma funcional del efecto de las covariables bajo el modelo de Cox clásico.

## 5.4. Aplicando el modelo de Cox con regresión splines

En la Tabla 5 se muestra los resultados del modelo de Cox con regresión splines (basado en un nodo y bases B-spline cúbico) para aproximar la forma funcional no lineal del efecto de las covariables continuas (edad, Hb, ln(leucocitos), linfocitos y ln(DHL)) en la función de riesgo, incluyéndose las covariables categóricas (género, zubrod, primario, estadio clínico y síntomas). La significancia del efecto no-lineal de las covariables contínuas se resume mediante la estadística de prueba local basado en test de Wald (contrasta los coeficientes de las bases B-splines).

Tabla 5. Resultados del modelo de Cox con regresión splines.

| Variables | $\hat{\beta}$ | EE($\hat{\beta}$) | $\chi^2$ (df) | p | HR (IC95 %) |
|---|---|---|---|---|---|
| Edad: | | | | | |
| — no-lineal | - | - | 16.55 (4) | 0.002 | bs ( , df=4) |
| Género masculino | 0.218 | 0.080 | 7.40 (1) | 0.007 | 1.24 (1.06, 1.45) |
| Zubrod 2-4 | 0.622 | 0.085 | 53.61 (1) | <0.001 | 1.86 (1.58, 2.20) |
| Primario ganglionar | -0.105 | 0.087 | 1.46 (1) | 0.227 | 0.90 (0.76, 1.07) |
| EC III-IV | 0.398 | 0.086 | 21.21 (1) | <0.001 | 1.49 (1.26, 1.76) |
| Síntomas B | 0.124 | 0.083 | 2.22 (1) | 0.136 | 1.13 (0.96, 1.33) |
| Hemoglobina: | | | | | |
| — no-lineal | - | - | 17.11 (4) | 0.002 | bs ( , df=4) |
| ln(leucocitos): | | | | | |
| — no-lineal | - | - | 12.91 (4) | 0.012 | bs ( , df=4) |
| Linfocitos: | | | | | |
| — no-lineal | - | - | 70.98 (4) | <0.001 | bs ( , df=4) |
| ln(DHL): | | | | | |
| — no lineal | - | - | 36.68 (4) | <0.001 | bs ( , df=4) |
| LR test (df) | | | 446.40 (25) | <0.001 | |
| Wald test (df) | | | 486.90 (25) | <0.001 | |
| AIC | 9531.92 | | | | |

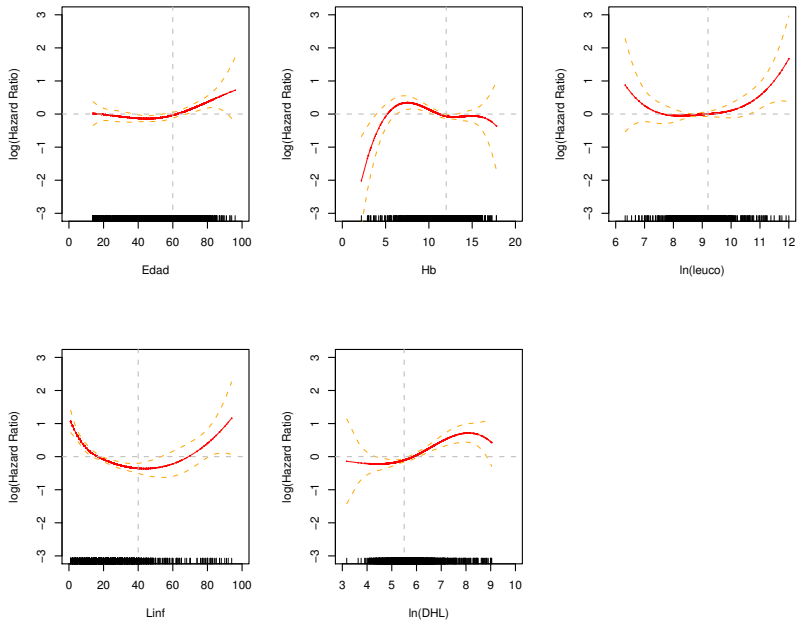
Nota: Las categorías que no aparecen corresponden a las categorías de referencia.
Beta: Coeficiente de regresión. EE: error estándar del coeficiente de regresión.
HR: Hazard ratio. bs: B-splines.

En los resultados (Tabla 5) se observa que las covariables incluidas en el modelo presentan un efecto significativo en la supervivencia global de los pacientes con LNH (test de LR y Wald: p<0.001, se rechaza la hipótesis $H_0 : \beta_1 = \ldots = \beta_p = 0$). Las covariables con efecto significativo ($p < 0.05$) fueron todas, a excepción del primario ($p = 0.227$) y los síntomas ($p = 0.136$).

Para las covariables categóricas la razón de riesgo de estas variables implica que, los pacientes de sexo masculino presentan un riesgo de mortalidad de HR=1.2 (IC5 %: 1.1-1.5) veces mas que los pacientes de sexo femenino. Los pacientes con zubrod 2-4 presentan un riesgo de mortalidad de HR=1.9 (IC95 %: 1.6-2.2) veces más que los pacientes con zubrod 0-1. Los pacientes con enfermedad avanzada (EC III-IV) presentan un riesgo de mortalidad de HR=1.5 (IC95 %: 1.3-1.8) veces más que los pacientes con enfermedad temprana (EC I-II).

Para las covariables continuas no se pueden realizar las mismas interpretaciones de la HR, debido a que el efecto de las covariables no presenta una relación lineal con la función de riesgo. En los resultados (Tabla 5) se observa que todas las variables continuas (edad, Hb, leucocitos, linfocitos y DHL) presentan un efecto no lineal significativo.

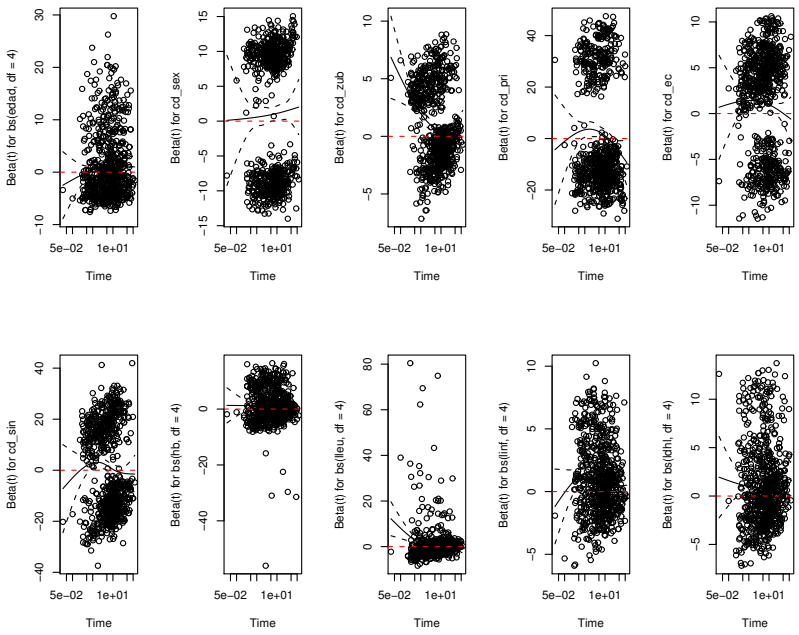
En la Gráfica 13 se muestra la forma funcional del efecto de las covariables continuas en el logaritmo de la razón de riesgo. En general, en todas estas covariables contínuas el riesgo es de forma no lineal. Para la variable edad, el riesgo de mortalidad para menores de 60 años es menor a HR=1, sin embargo, el riesgo se incrementa después de los 60 años de edad. Para la variable hemoglobina (Hb), el riesgo de mortalidad para Hb >12g/dl es menor a HR=1, sin embargo, el riesgo se incrementa a medida que el nivel de hemoglobina disminuye después de 12g/dl. Para el número de leucocitos, el riesgo de mortalidad se incrementa a medida que el número de leucocitos disminuye después de 3 mil (efecto de leucopenia) o se incrementa después de 10 mil leucocitos (efecto de leucocitosis).




Gráfica 13. Forma funcional del efecto de las covariables mediante el modelo de Cox con regresión splines.

Para el porcentaje de linfocitos, el riesgo de mortalidad se incrementa cuando el porcentaje de linfocitos disminuye después de 20 % o se incrementan después de 60 %. Para la DHL, el riesgo de mortalidad se incrementa después de 240 $UI/L$.

En la Gráfica 14 se muestra los residuos de Schoenfeld escalados vs. tiempo. En los gráficos se observa que el supuesto de riesgos proporcionales no se verifican para las siguientes variables: zubrod, primario y los linfocitos. Los cuales, se verifican formalmente mediente el test de no-proporcionalidad.



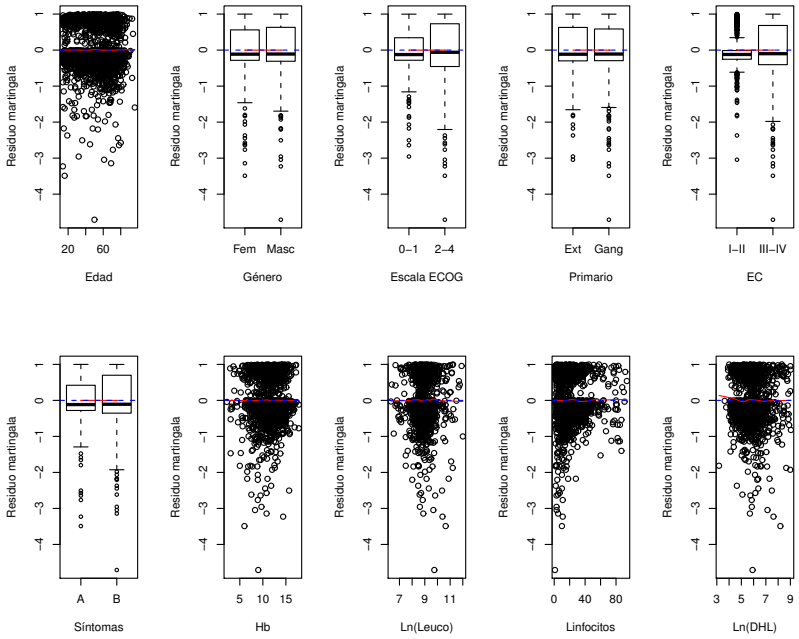

Gráfica 14. Residuos de Schoenfeld escalados del modelo de Cox con regresión splines. De izquierda a derecha, en la primera fila se muestra para la edad, género, zubrod, primario, estadio clíncio, y en la segunda fila para síntomas, Hb, ln(leucocitos), linfocitos y ln(DHL)

En la Tabla 6 se muestran los resultados del test de no proporcionalidad de Therneau y Grambsch. En los resultados se observan que el efecto de las covariables no verifican el supuesto de riesgos proporcionales (riesgos constantes) para las siguientes variables: zubrod, primario y linfocitos ($p > 0.05$).

En la Gráfica 15 se muestran los residuos martingala para el modelo de Cox con regresión splines. En los gráficos se observa que los residuos son constantes (líneas rojas) en relación a cada covariable, lo cual verifica que los efectos de las covariables son bien aproximados por bases B-spline cúbico.

TABLA 6. Test de no proporcionalidad de Therneau y Grambsch, basado en los residuos de Schoenfeld escalados del modelo de Cox con regresión splines.

| Variables | correlación de Pearson (rho) | $\chi^2$ | valor-p |
|---|---|---|---|
| bs(edad, df = 4) | 0.0610 | 2.632 | 0.105 |
| Género masculino | 0.0374 | 1.014 | 0.314 |
| Zubrod 2-4 | -0.1958 | 26.121 | <0.001 |
| Primario ganglionar | -0.1000 | 7.592 | 0.006 |
| Estadio clínico III-IV | -0.0659 | 3.276 | 0.070 |
| Síntomas B | -0.0677 | 3.316 | 0.069 |
| bs(hb, df = 4) | -0.0457 | 1.399 | 0.237 |
| bs(lleu, df = 4) | -0.0694 | 3.538 | 0.060 |
| bs(linf, df = 4) | -0.0975 | 6.415 | 0.011 |
| bs(ldhl, df = 4) | 0.0140 | 0.144 | 0.704 |
| Global | NA | 93.785 | <0.001 |




GRÁFICA 15. Residuos martingala del modelo de Cox con regresión splines.

En conclusión, si bien el modelo de Cox con regresión splines (basado en un nodo y bases B-splines cúbico) aproxima bién la forma funcional del efecto de las covariables en el logaritmo de la razón de riesgo, estas aún presentan problemas de no cumplimiento del supuesto de riesgos proporcionales para algunas covariables; por tanto, los resultados de las estimaciones bajo el modelo son aún discutible para los datos de los pacientes con LNH.

## 5.5. Aplicando el modelo de Cox con P-splines

En la Tabla 7 se muestra los resultados del modelo de Cox con P-splines (basado en 10 nodos y bases B-spline cúbico) para aproximar la forma funcional no lineal del efecto de las covariables continuas (edad, Hb, ln(leuco), linfocitos y ln(DHL)) en la función de riesgo, incluyéndose las covariables categóricas (género, zubrod, primario, estadio clínico y síntomas).

TABLA 7. Resultados del modelo de Cox con P-splines.

| Variables | $\hat{\beta}$ | EE($\hat{\beta}$) | $\chi^2$ (df) | p | HR (IC95 %) |
|---|---|---|---|---|---|
| Edad: | | | | | |
| — lineal | 0.006 | 0.002 | 7.15 (1) | 0.007 | |
| — no-lineal | - | - | 9.25 (3) | 0.028 | P-splines ( , df=4) |
| Género masculino | 0.210 | 0.080 | 6.84 (1) | 0.009 | 1.23 (1.05, 1.44) |
| Zubrod 2-4 | 0.620 | 0.085 | 52.99 (1) | <0.001 | 1.86 (1.57, 2.20) |
| Primario ganglionar | -0.113 | 0.086 | 1.72 (1) | 0.190 | 0.89 (0.75, 1.06) |
| EC III-IV | 0.397 | 0.096 | 21.13 (1) | <0.001 | 1.49 (1.26, 1.76) |
| Síntomas B | 0.122 | 0.083 | 2.18 (1) | 0.140 | 1.13 (0.96, 1.33) |
| Hemoglobina: | | | | | |
| — lineal | -0.031 | 0.018 | 3.03 (1) | 0.082 | |
| — no-lineal | - | - | 12.61 (3) | 0.006 | p-splines ( , df=4) |
| ln(leucocitos): | | | | | |
| — lineal | 0.110 | 0.063 | 3.02 (1) | 0.082 | |
| — no-lineal | - | - | 8.45 (3) | 0.038 | p-splines ( , df=4) |
| Linfocitos: | | | | | |
| — lineal | -0.011 | 0.003 | 16.71 (1) | <0.001 | |
| — no-lineal | - | - | 49.70 (3) | <0.001 | p-splines ( , df=4) |
| ln(DHL): | | | | | |
| — lineal | 0.254 | 0.045 | 31.52 (1) | <0.001 | |
| — no lineal | - | - | 6.60 (3) | 0.089 | p-splines ( , df=4) |
| LR test (df) | | | 455.00 (25.2) | <0.001 | |
| Wald test (df) | | | 473.00 (25.2) | <0.001 | |
| AIC | 9523.97 | | | | |

Nota: Las categorías que no aparecen corresponden a las categorías de referencia.
Beta: Coeficiente de regresión. EE: error estándar del coeficiente de regresión.
HR: Hazard ratio.

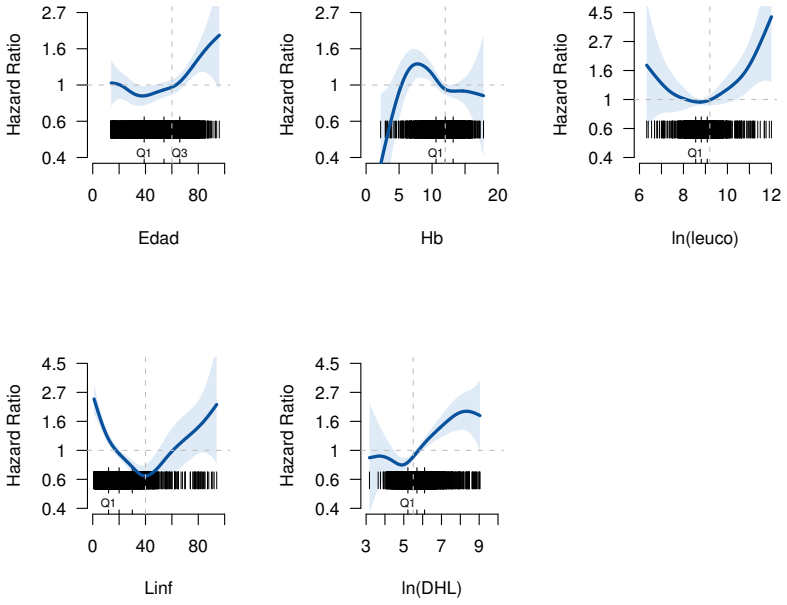
En los resultados (Tabla 7) se observa que las covariables presentan un efecto significativo en la supervivencia global de los pacientes con LNH (test de LR y Wald: p<0.001, se rechaza la hipótesis $H_0 : \beta_1 = \ldots = \beta_p = 0$). Las covariables con efecto significativo ($p < 0.05$)

fueron todas, a excepción del primario ($p < 0.190$) y síntomas ($p < 0.140$), tal como fue identificado con regresión splines.

Para las covariables categóricas la razón de riesgo de estas variables implica que, los pacientes de sexo masculino presentan un riesgo de mortalidad de HR=1.2 (IC5 %: 1.1-1.4) veces mas que los pacientes de sexo femenino. Los pacientes con zubrod 2-4 presentan un riesgo de mortalidad de HR=1.9 (IC95 %: 1.6-2.2) veces más que los pacientes con zubrod 0-1. Los pacientes con enfermedad avanzada (EC III-IV) presentan un riesgo de mortalidad de HR=1.5 (IC95 %: 1.3-1.8) veces más que los pacientes con enfermedad temprana (EC I-II).

Para las covariables continuas no se pueden realizar las mismas interpretaciones de la HR, debido a que el efecto de las covariables no presenta una relación lineal con la función de riesgo. En los resultados (Tabla 7) se observa que todas las variables continuas (edad, Hb, leucocitos, linfocitos y DHL) presentan un efecto no lineal significativo.

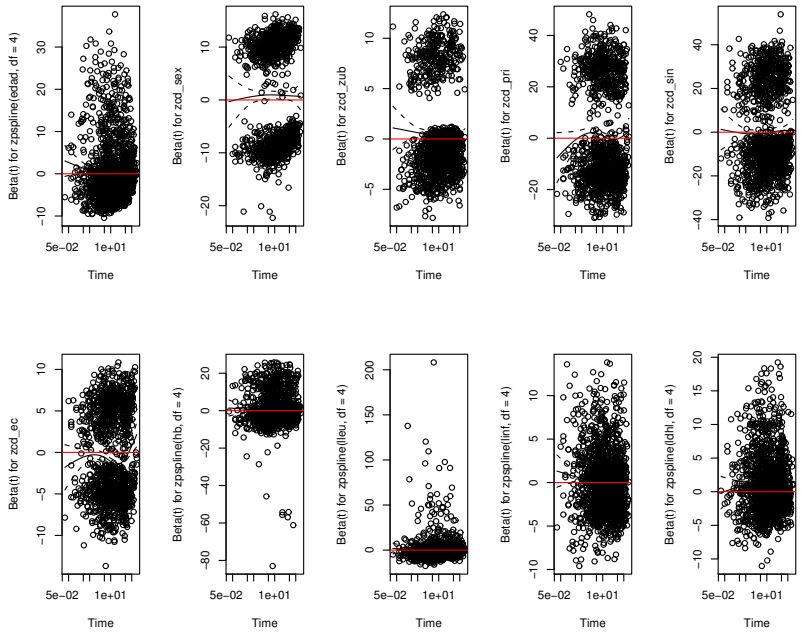
En la Gráfica 16 se muestra la forma funcional del efecto de las covariables continuas en la razón de riesgo. Para la variable edad, el riesgo de mortalidad para menores de 60 años es menor a HR=1, sin embargo, el riesgo se incrementa después de los 60 años de edad en una forma no lineal. Para la variable hemoglobina (Hb), el riesgo de mortalidad para Hb >12g/dl es menor a HR=1, sin embargo, el riesgo se incrementa a medida que el nivel de hemoglobina disminuye después de 12g/dl.




Gráfica 16. Forma funcional del efecto de las covariables mediante el modelo de Cox con P-splines.

Para el número de leucocitos, el riesgo de mortalidad se incrementa a medida que el número de leucocitos disminuye después de 3 mil (efecto de leucopenia) o se incrementa después de 10 mil leucocitos (efecto de leucocitosis). Para el porcentaje de linfocitos, el riesgo de mortalidad se incrementa cuando el porcentaje de linfocitos disminuye después 20 % o se incrementan después de los 60 %. Para la DHL, el riesgo de mortalidad se incrementa después de 240 $UI/L$.

En la Gráfica 17 se muestra los residuos de Schoenfeld escalados vs. tiempo. En los cuales, se observa que las curvas de ajuste se aproximan a una línea recta horizontal, lo que sugiere el cumplimiento del supuesto de riesgos proporcionales.



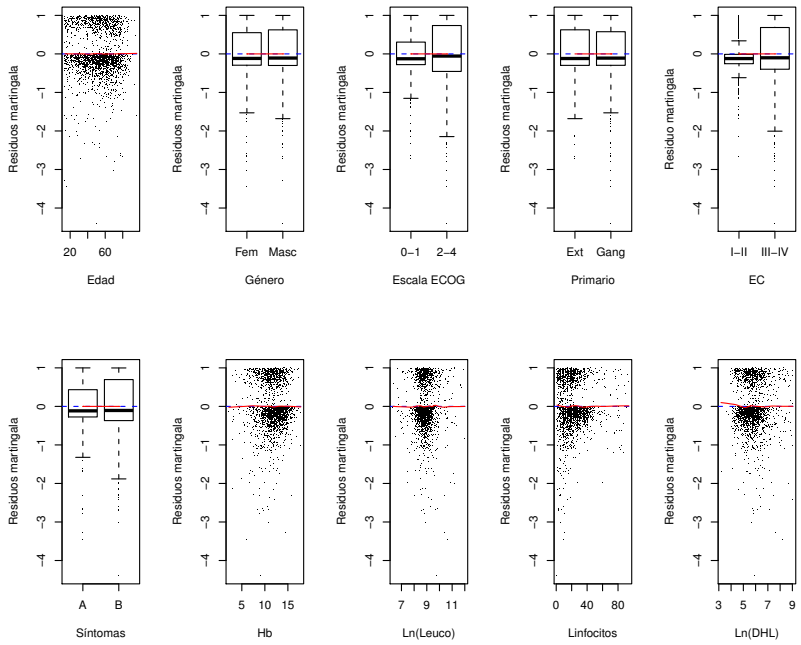

GRÁFICA 17. Residuos de Schoenfeld escalados vs. tiempo ($ln(meses)$) del modelo de Cox con P-splines. De izquierda a derecha, en la primera fila muestra para la edad, género, zubrod, primario, estadio clíncio, y en la segunda fila para síntomas, Hb, ln(leucocitos), linfocitos y ln(DHL)

En la Tabla 8 se muestran los resultados del test de no proporcionalidad de Therneau y Grambsch. En los resultados se observan que el efecto de las covariables cumplen el supuesto de riesgos proporcionales (riesgos constantes) ($p > 0.05$). Por lo tanto, los resultados de las estimaciones bajo el modelo de Cox con P-splines son válidos.

En la Gráfica 18 se muestran los residuos martingala para el modelo de Cox con P-splines. Los resultados indican que los residuos son constantes y se aproximan a una recta horizontal; lo cual verifica que el efecto de las covariables son bien aproximados por los P-splines.

TABLA 8. Test de no proporcionalidad de Therneau y Grambsch, basado en los residuos de Schoenfeld escalados del modelo de Cox con P-splines.

| Variables | correlación de Pearson (rho) | $\chi^2$ | valor-p |
|---|---|---|---|
| pspline(edad, df = 4) | 0.01664 | 0.43494 | 0.5096 |
| Género masculino | 0.00123 | 0.00220 | 0.9626 |
| Zubrod 2-4 | -0.05522 | 4.48729 | 0.0341 |
| Primario ganglionar | -0.01379 | 0.28662 | 0.5924 |
| Síntomas B | 0.02555 | 0.98132 | 0.3219 |
| Estadio clínico III-IV | 0.01355 | 0.26061 | 0.6097 |
| pspline(hb, df = 4) | -0.00132 | 0.00263 | 0.9591 |
| pspline(lleu, df = 4) | 0.00780 | 0.10383 | 0.7473 |
| pspline(linf, df = 4) | -0.05655 | 4.63394 | 0.0313 |
| pspline(ldhl, df = 4) | -0.02377 | 0.72981 | 0.3929 |
| Global | NA | 13.89722 | 0.1777 |




GRÁFICA 18. Residuos martingala del modelo de Cox con P-splines.

En conclusión, el modelo de Cox con P-splines (basado en 10 nodo y bases B-splines cúbico) aproxima bién la forma funcional del efecto de las covariables. Si bien la edad, zubrod, estadio clínico y la DHL son factores pronósticos muy conocidos en esta patología, aquí además se identifica la hemoglobina, el número de leucocitos y el porcentaje de linfocitos con efectos significativos ($p < 0.05$).

## 5.6. Comparación de los modelos

En la Tabla 9 se muestra un resúmen de los resultados obtenidos con los tres modelos (modelo de Cox clásico, modelo de Cox con regresión splines y modelo de Cox con P-splines). Aquí se comparan los factores pronósticos identificados, la forma funcional del efecto de las covariables en la función de riesgo y la selección del mejor modelo mediante AIC (criterio de información de Akaike)

En el modelo de Cox clásico, los factores pronósticos con efecto significativo para la supervivencia global de los pacientes con LNH fueron todas a excepción del foco primario (p=0.580). Sin embargo, en este modelo las variables como zubrod, primario, síntomas y ln(leucocitos) no cumplieron el supuesto de riesgos proporcionales; así mismo, los residuos martingala no verificarón el supuesto de relación lineal entre el logaritmo de la razón de riesgo y las covariables contínuas. Estos resultados sugerieron utilizar métodos más flexibles para aproximar el efecto no-lineal de las covariables contínuas.

En cambio con el modelo de Cox con regresión splines y modelo de Cox con P-splines las covariables con efecto significativo fueron todas a excepción del primario y síntomas ($p > 0.05$). Los residuos martingala de ambos modelos en relación a las covariables contínuas no muestran un patrón que sugiera la existencia de alguna forma funcional que evidencie la falta de ajuste de los datos. Para ambos modelos los residuos martingala en relación a las covariables contínuas son aproximadamente constantes, en consecuencia el efecto no lineal de las covariables contínuas es bién aproximado por ambos modelos.

Sin embargo, si ambos modelos describen bién la forma funcional no lineal del efecto de las covariables, el ajuste bajo el modelo de Cox con P-splines verifica el supuesto de riesgos proporcionales (riesgos constantes), que es una suposición fuerte del modelo de Cox, en cambio el ajuste mediante el modelo de Cox con regresión splines no cumple el supuesto de riesgos proporcionales.

Por otro lado, de acuerdo al criterio de información de Akaike (AIC), el modelo de Cox con P-splines presenta ligeramente una menor AIC (9523.97) en comparación al modelo de Cox con regresión splines (AIC: 9531.92).

En la Gráfica 19 se muestra las curvas de supervivencia estimada para cada modelo. En los resultados se observa que las curvas de supervivencia estimada para los tres modelos son muy similares y están dentro de las bandas de confianza de la curva de supervivencia estimada mediante el método de Kaplan-Meier.

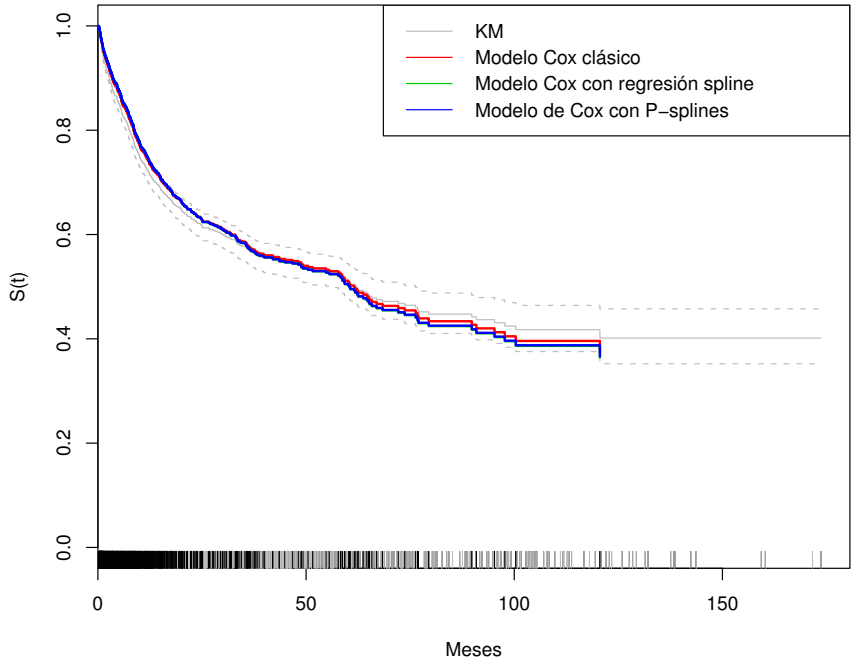
TABLA 9. Resultados del modelo de Cox con regresión splines y del modelo de Cox con P-splines.

| Variables | Modelo de Cox | | Modelo de Cox con RS | | | Modelo de Cox con PS | | |
|---|---|---|---|---|---|---|---|---|
| | $p$ | HR | R-spline | $p$ | HR | P-splines | $p$ | HR |
| Edad: | 0.027 | 1.01 | bs(,df=4) | 0.002 | no-lineal | pspline(, df=4) | 0.028 | no-lineal |
| Género masculino | 0.010 | 1.22 | - | 0.007 | 1.24 | - | 0.009 | 1.23 |
| Zubrod 2-4 | <0.001 | 1.99 | - | <0.001 | 1.86 | - | <0.001 | 1.86 |
| Primario gangl. | 0.580 | 0.96 | - | 0.227 | 0.90 | - | 0.109 | 0.89 |
| EC III-IV | <0.001 | 1.55 | - | <0.001 | 1.49 | - | <0.001 | 1.49 |
| Síntomas B | 0.045 | 1.18 | - | 0.136 | 1.13 | - | 0.140 | 1.13 |
| Hb: | 0.002 | 0.95 | bs(,df=4) | 0.002 | no-lineal | pspline(, df=4) | 0.006 | no-lineal |
| ln(Leucocitos): | 0.002 | 1.24 | bs(,df=4) | 0.012 | no-lineal | pspline(, df=4) | 0.038 | no-lineal |
| Linfocitos: | <0.001 | 0.99 | bs(,df=4) | <0.001 | no-lineal | pspline(, df=4) | <0.001 | no-lineal |
| ln(DHL): | <0.001 | 1.29 | bs(,df=4) | <0.001 | no-lineal | pspline(, df=4) | 0.089 | no-lineal |
| LR test | 351.00 | | | 446.40 | | | 455.00 | |
| Wald test | 379.10 | | | 486.90 | | | 473.00 | |
| AIC | 9597.33 | | | 9531.92 | | | 9523.97 | |

Nota: RS: regresión splines, PS: P-splines y $ln$: logaritmo natural.

La Tabla 10 se muestra los factores pronósticos para la supervivencia global de los pacientes con LNH, bajo el modelo de Cox con P-splines. Las covariables con efecto significativo a un nivel de significación de 5 % fueron: la edad (p = 0.028), el género (p = 0.009), zubrod (p < 0.001), estadio clínico (p < 0.001), hemoglobina (p = 0.006), número de leucocitos (p = 0.038), porcentaje de linfocitos (p < 0.001), así como el efecto de la DHL (p = 0.089) por ser clínicamente relevante en los LNH. La curva de HR para estas covariables se muestra en la Gráfica 16.

.




GRÁFICA 19. Curvas de supervivencia estimada de acuerdo al modelo de Cox: clásico, con regresión splines y P-splines.

TABLA 10. Factores pronósticos para a supervivencia global de los pacientes con LNH, bajo el modelo de Cox con P-spline.

| Variables | p | HR (IC95 %) |
|---|---|---|
| Edad | 0.028 | p-splines (, df=4) |
| Género masculino | 0.009 | 1.23 (1.05, 1.44) |
| Zubrod 2-4 | <0.001 | 1.86 (1.57, 2.20) |
| Primario ganglionar | 0.190 | 0.89 (0.75, 1.06) |
| EC III-IV | <0.001 | 1.49 (1.26, 1.76) |
| Síntomas B | 0.140 | 1.13 (0.96, 1.33) |
| Hb | 0.006 | p-splines (, df=4) |
| ln(leuco) | 0.038 | p-splines (, df=4) |
| Linfocitos | <0.001 | p-splines (, df=4) |
| ln(DHL) | 0.089 | p-splines (, df=4) |

Nota: Las categorías que no aparecen corresponden a las categorías de referencia.
HR: Hazard ratio. IC95 %: Intervalo de confianza al 95 %.

# Capítulo 6
# Discusión y conclusiones

El amplio uso de los modelos tradicionales para el análisis de datos de supervivencia y el desarrollo de programas computacionales han contribuido al desarrollo de métodos más sofisticados de análisis de supervivencia desde técnicas simples a más complejas, las cuales han crecido rápidamente durante los últimos años para un mejor modelamiento, facilitado por el desarrollo de la tecnología computacional.

Si bien el modelo de Cox (1972) es una herramienta muy utilizada para determinar el efecto de las covariables en muchos contextos estadísticos, este modelo está sujeto al cumplimiento de los supuestos como son: riesgos proporcionales, covariables invariantes en el tiempo y que la estructura de la relación entre la función de riesgo y las covariables sea lineal. Sin embargo, estas condiciones o restricciones no necesariamente se cumplen en muchas aplicaciones. En este sentido, la no-linealidad puede ser tan frecuente como el no cumplimiento de riesgos proporcionales; como algunos autores refieren uno puede ser consecuencia del otro, es decir, si no hay proporcionalidad es muy posible que tampoco haya linealidad (Keele, 2010).

En consecuencia, si el supuesto de riesgos proporcionales no se cumplen, los resultados bajo el modelo de Cox clásico no es el más adecuado, entonces el modelo de Cox estratificado, modelo de Cox extendido con variable tiempo-dependiente, modelo de Cox ponderado, modelo de odds proporcional o el modelo log-logístico podrían ser una alternativa; sin embargo, en todos estos modelos se deberá tener en cuenta la forma lineal del efecto de las covariables contínuas.

En este trabajo, se utilizaron métodos más flexibles como son: regresión splines y suaviamiento splines penalizado (P-spline), debido a que en nuestros datos, la forma funcional no satisface el supuesto de relación lineal en el modelo de Cox clásico. En cambio utilizando el modelo de Cox con regresión splines y el modelo de Cox con P-splines se obtuvieron una mejor aproximación de los efectos de las covariables en la función de riesgo. En consecuencia, la razón de riesgo para cada covariable contínua presenta una estructura de relación cuya forma funcional es no lineal.

Los factores pronósticos con efecto significativo para la supervivencia en LNH fueron: edad, género, zubrod, estadio clínico, nivel de hemoglobina, leucocitos, linfocitos y la DHL mediante el modelo de Cox con P-splines, así como através del modelo de Cox con regresión splines (aunque en este último modelo no se pudo verificar el supuesto de riesgos proporcionales para este método de suavizamiento). Los cuales, concuerdan con los reportados en la literatura para esta patología (Nicolaides, Dimos y Pavlidis, 1998; Rebasa, 2001).

Cabe resaltar que los puntos de corte (HR=1) determinados para las covariables continuas mediante estos métodos se aproximan a los puntos de corte definidos clínicamente como grupos de peor pronóstico para algunas de las variables continuas. Según los resultados del modelo de Cox con P-splines, los pacientes mayores de los 60 años de edad tienen un peor pronóstico, el cual coincide con el punto de corte definido para clasificar a los pacientes según la edad en grupo de peor pronóstico. Para la Hb baja ($<12g/dl$) y los valores elevados de la DHL ($>240U/L$) los puntos de corte también coinciden con los puntos de corte definidos clínicamente para grupos de peor pronóstico.

Sin embargo, para los valores de los leucocitos y los linfocitos existen dos puntos de corte claramente definidos que muestran un mayor riesgo de mortalidad: i) leucocitos menores de $3mil$ y mayores de $10mil$, y ii) linfocitos menores de 20 % y >60 %. Estos grupos de riesgo deberían ser considerados en la práctica clínica al momento de clasificar a los pacientes en grupos de pronóstico y así optimizar el benficio del tratamiento.

Finalmente algunas limitaciones de ste trabajo están referidas en cuanto a la base de datos disponible para realizar el análisis. Todos los datos fueron recopilados retrospectivamente de las historias clínicas de los pacientes; en las cuales la mayoría de los datos no fueron registrados de acuerdo a los objetivos de este estudio. Resultado de esto son los diferentes criterios de clasifcación histopatológica que no han permitido incluir en el análisis algunas variables como tipo histológico, grados de agresividad e inmunofenotivo (tipo celular), así como las $\beta$-2 microglobulinas que son factores pronósticos en este grupo de pacientes.

Recomendaciones para trabajos posteriores:

- Desde el aspecto clínico, se podría plantear la necesidad de realizar el análisis de los factores pronósticos en los LNH agresivos, principalmente linfomas de células grandes B difuso, que según la clasifcación de la Organización Mundial de la Salud representa casi el 80 % de los LNH, y verificar el valor pronóstico del IPI (Shipp et al. (1993)) y su capacidad predictiva de acuerdo a los nuevos métodos estadísticos que son más flexibles, y proponer otro índice de pronóstico incluyendo variables como $\beta - 2$ microglobulinas que está muy relacionado al crecimiento tumoral.

Otro aspecto importante en este grupo de pacientes sería determinar los factores pronósticos con métodos flexibles en pacientes con LNH de céluas T, que actualmente presentan un peor pronóstico que los LNH de células B.

- Desde el aspecto metodológico, se podría plantear la necesidad de realizar el análisis de los factores pronósticos utilizando el modelo de Cox con splines penalizados (P-splines) para aproximar la forma funcional no-lineal del efecto de las covariables junto con los coeficientes tiempo-dependiente no-lineales para las covariables que no son constantes en el tiempo (Abrahamowicz et al. (1996); Abrahamowicz et al. (2007)).
  Otros aspecto metodológico importante en el modelo de Cox con métodos flexibles serían: estimar la función de riesgo base junto con los efectos de las covariables de manera simultánea y no como se realiza en el modelo de Cox clásico, valorar el efecto espacial y temporal en la supervivencia de los pacientes con LNH, sabiendo que la sobrevida de los pacientes pueden variar de acuerdo a la distribución geográfica (efecto espacial), sea departamental o regional, por lugar de nacimiento o procedencia, y por periodos de tiempo (period analysis), así como el efecto de las variables intermedias durante cada periodo tratamiento y seguimiento (multi-state models).

# Bibliografía


[1] Abrahamowicz, M., MacKenzie T. and Esdaile, J.M. (1996). Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. JASA, Vol. 91, pp. 1432-1439.

[2] Abrahamowicz, M. and MacKenzie, T. (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. Statistics in Medicine, Vol. 26, pp. 392-408.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Inference Theory, B.N. Petrov and F. Csáki (eds), pp. 267-281. Budapest: Akadémiai Kiadó.

[4] Ata, N. and Tekin M. (2007) Cox regression models with non-proportional hazard applied to lung cancer survival data. Journal of Mathematics and Statistics, Vol. 36, pp. 157-167.

[5] Arece, F. y Rodríguez, D. (2003) Linfoma no Hodgkin agresivo: ¿Después del CHOP sólo el CHOP?. Rev. Cubana Med, 42(1), pp. 79-88.

[6] Ambler, G. and and Royston, P. (2001) Fractional polynomial model selection procedure: Investigation of type I error rate. Journal of Statistical Computation and Simulation, Vol. 69, pp. 89-108.

[7] Costas, N., Dimou, S., Pavlidis, N. (1998) Prognostic Factors in Agressive non-Hodgkins lymphomas The Oncologia, Vol. 3, pp. 189-197

[8] Cox, D.R. (1972) Regression models and life tables (with discussion), Journal of the Royal Statistical Society, Vol. 34, pp. 187-220

[9] Dabrowska D.M. (1997) Smoothed Cox model. The Annals of Statistics, Vol. 25, pp. 1510-1540.

[10] De Boore, C. (1977). Package for calculating B-splines. J. Numer. Anal., Vol. 14, pp. 441-472.

[11] Dierckx, P. (1993). Curve and Surface Fitting with Splines, Clarendon, Oxford (UK).

[12] Eilers, P.H.C. and Marx, B.D. (1996) "Flexible smoothing with B-splines and penalties". Statistical Science, 1996, Vol. 11, pp. 98-102.

[13] Eubank, R.L. (1988) Spline smoothing and nonparametric regression. Marcel Dekker, New York.

[14] Friedman, J.H. (1991) Multivariate adaptive regression splines. The Annals of Statistics, Vol. 19, pp. 1-67.

[15] Friedberg, J.W., Mauch, P.M., Rimsza, L.M. and Fisher, R.I. (2008) Non-Hodgkin's lymphomas. In: DeVita VT, Lawrence TS, Rosenberg SA, eds. DeVita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology. 8th ed. Philadelphia, Pa: Lippincott Williams & Wilkins; pp. 2278-2292.

[16] Gray, R.J. (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. JASA, Vol. 87, pp. 942-951.

[17] Green, P.J. and Silverman, B.W. (1994) Nonparametric regression and generalized linear models. Chapman & Hall, New York.

[18] Globocan 2002: IARC. Lyon - Francia. (http://www-dep.iarc.fr/)

[19] Hastie, T.J. and Tibshirani R.J. (1990) Generalized additive models. London: Chapman and Hall.

[20] Hartge, P. and Smith, M.T. (2007) Environmental and behavioral factors and the risk of non-Hodgkin lymphoma. Cancer Epidemiol Biomarkers Prev. Vol. 16, pp. 367-368.

[21] Horsman, J.M. and Hancock, H. (2001) Prognostic Markers in Malignant Lymphoma. An Analysisi of 1198 Patiente treated at a single centre. International Journal of Oncology, Vol. 19, pp. 1203-1209.

[22] Keele, L. (2010) Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models. Political Analysis, Vol. 18, pp. 189-205.

[23] Kyle, F. and Hill, M. (2010) NHL (diffuse large B-cell lymphoma). Clinical Evidence, 2010;11:2401

[24] Klein, J.P. and Moeschberger, M.L. (1997) Survival analysis: Technique for censored and truncated data. Springer-Verlag, New York.

[25] Lin, D.Y. and Wei, L.J.Z. (1993) Checking the Cox model with cumulative sums of martingale based residuals. Biometrics, Vol. 80, pp. 557-572.

[26] Martinussen, T. and Scheike, T. (2006) Dynamic regression models for survival data. Springer, New York.

[27] Mounier, N., Diviné, M., Haioun, C., Lepage, E. and Reyes, F. (1997) Factores pronósticos de los linfomas malignos. J. García-Conde, E. Matutes, M.A. Piris, F. Reyes editores. Síndromes Linfoproliferativos. Productos Roche S.A. 1ra edición. pp. 69-77.

[28] Muir, C., Waterhaus, J., Mack, T. Powell, J. and Whelan, S. (1987) Cancer Incidence in Five Continents. Vol. V, IARC Scientific Publication n° 88, Lyon.

[29] Nicolaides, C., Dimous, S. and Pavlidis, N. (1997) Prognostic factor in aggressive non-Hodgkin lymphomas. The Oncologist, Vol.3, pp.189-197.

[30] O'Sullivan, F. (1988) "Nonparametric estimation of relative risk functions using splines and cross-validation."SIAM Journal on Scientific and Statistical Computing, Vol.9, pp. 531-542.

[31] Parkin, D.M., Whelan, S.L., Ferlay, J., Teppo, L. and Thomas D.B. (2002) Cancer Incidence in Five Continents. Vol. VIII, IARC Scientific Publications n° 145, Lyon.

[32] Programas Nacionales de Control de Cáncer: Políticas y Pautas para la Gestión. OPS, 2004.

[33] Rabasa, MP. (2001) Factores pronósticos en los linfoma no Hodgkin y linfoma de Hodgkin. ANNALES Sis San Navarra, Vol.24 (Supl. 1), pp. 141-158.

[34] Royston, P. and Sauerbrei, W. (2008) Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley.

[35] Royston, P. and Altman, D.G. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Appl. Statist. Vol43, pp. 429-467.

[36] Sauerbrei, W. and Royston, P. (1999): Building multivariable prognosis and diagnostic models: transformation of the predictors by using fractional polynomials. J. Roy. Statist. Soc. Ser. A. Vol. 162, pp. 71-94.

[37] Shipp M.A., Harrington D.P., Aderson J.R. et al. (1993). A predictive model for aggressive non-Hodgkin's lymphoma. The International non-Hodgkin's lymphoma prognostic factors project. *N Engl J Med* 329: 987-994.

[38] Sleeper, L.A and Harrington, D.P. (1990) Regression splines in the Cox model with application to covariate effects in liver disease. JASA, 1990, Vol. 85, pp. 941-949.

[39] Stone, C.J. (1985) Additive regression and other nonparametric models. The Annals of Statistics, Vol.13, pp. 689-705.

[40] Terje, K.J. (undated) Nonparametric estimation in Cox-model: time transformation methods versus partial likelihood methods.

[41] Therneau, T.M. and Grambsch, P.M. (2000) Modeling survival data: extending the Cox model. Springer-Verlag, New York.

[42] Therneau, T.M., Grambsch, P.M. and Fleming T.R. (1990) Martingale-based residuals for survival data. Biometrika, Vol. 77, pp. 147-160.

[43] Verweij, P.J.M. and van Houwelingen, H.C. (1993) Crossvalidation in survival analysis. Statistics in Medicine, Vol. 12, pp. 2305-2314.

[44] Wang, W., Wang, J.L. and Wang, Q. (2009) Proportional hazards regression with unknown link function. IMS Lecture Notes-Monograph Series Optimality The Third Erich L. Lehmann Symposium, Vol. 57, pp. 47-66.