TRABAJO DE FIN DE MÁSTER

Máster en Estadística e Investigación Operativa

Título: Modelo de Regresión de Cox con Métodos Flexibles en Pacientes con Linfoma No Hodgkin

Autor: Claudio Jaime Flores Flores

Director: Guadalupe Gómez Melis

Departamento: Departamento de Estadística e Investigación Operativa

Convocatoria: 31 / enero / 2011

Facultat de Matemàtiques i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Tesis de Master

MODELO DE REGRESIÓN DE COX CON METODOS FLEXIBLE EN PACIENTES CON LINFOMA NO HODGKIN

Claudio Jaime Flores Flores

Director: Guadalupe Gómez Melis

Departamento de Estadística e Investigación Operativa

A mis padres: Alejandro (en memoría) y Brígida

Resumen

En muchos estudios clínicos es muy frecuente el uso de modelo de riesgo proporcional 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 una aplicación de los métodos flexibles como P-splines y polinomio fraccional para determinar y aproximar la forma funcional del efecto de las covariables (factores pronósticos) en la supervivencia global de los pacientes con LNH. Los resultados muestran que el efecto de las covariables continuas como Hb, Leucocitos, linfocitos y la DHL presentan una forma funcional no lineal con el logaritmo de la razón de riesgo.

Palabras clave: Modelo de Cox, P-spline, Polinomio Fraccional, LNH.

MSC2000: Codis de la *Mathematic Subject Classification*

Abstract

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 an application of flexible methods such as P-spline and fractional polynomial to identify, and align the functional form of the effect of covariates (prognostic factors) in the overall survival of patients with NHL. These results show that the effect of continuous covariates as: Hb, leukocytes, lymphocytes and LDH have a nonlinear functional form with the log hazard ratio.

Keywords: Cox model, P-spline, polynomial fractional, NHL.

MSC2000: 2000Mathematical Subject Classification

Índice general

Capítulo 1. Introducción 1
1.1. Introducción 1
1.2. Objetivos del trabajo 2
1.3. Detalles del desarrollo 3

Capítulo 2. LNH: Aspectos clínicos y metodológicos 5
2.1. Linfoma No Hodgkin 5
2.2. Supervivencia y pronóstico 6
2.3. Factores pronósticos 7
2.4. Descripción de los datos 8

Capítulo 3. Conceptos básicos en análisis de supervivencia 11
3.1. Datos en análisis de supervivencia 11
3.2. Funciones del tiempo de supervivencia 14
3.3. Función de máxima verosimilitud 15
3.4. Procesos de conteo 16
3.5. Suavizamiento con splines 19

Capítulo 4. Modelo de regresión de Cox con métodos flexibles 27
4.1. Revisión de la literatura 27
4.2. Modelo de regresión de Cox 29
4.3. Modelo de regresión de Cox con P-splines 33
4.4. Modelo de regresión de Cox con polinomio fraccional 38
4.5. Métodos de diagnóstico en el modelo de Cox 39

Capítulo 5. Factores pronósticos en LNH 47
5.1. Descripción de los datos 47
5.2. Aplicando el modelo de Cox clásico 48
5.3. Aplicando el modelo de Cox con P-splines 52
5.4. Aplicando el modelo de Cox con PF 56
5.5. Comparación de los modelos 58

Capítulo 6. Discusión y conclusiones 63

Bibliografía 65

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.

En general, los tiempos de supervivencia son muestras incompletas o censuradas, por lo que la técnica estadística relacionada con este tipo de datos se denomina análisis de supervivencia en el área de salud, y otras denominaciones en otras áreas como la ingeniería, economía y ciencias sociales.

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

En muchos estudios clínicos el objetivo del análisis puede ser evaluar el efecto de las covariables en la supervivencia de los pacientes. 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 los resultados de las estimaciones bajo el modelo 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 los métodos de suavizamiento como los splines y el método de polinomio fraccional.

Si bien el modelo de riesgos proporcionales de Cox forma parte de la asignatura de Análisis de Supervivencia que se imparte en el Máster Interuniversitario en Estadística e Investigación Operativa en la Facultad de Matemáticas y Estadística de la Universidad Politécnica de Cataluña, y sirve de base teórica, ya que el curso no cubre las extensiones del modelo de Cox a otras aproximaciones para modelar el efecto de las covariables en la función de riesgo de una manera más flexible utilizando los splines y los polinomios fraccionales.

En este trabajo se analizan el efecto de las covariables en la supervivencia de los pacientes con linfoma no Hodgkin (LNH), utilizando estos dos métodos (P-splines y Polinomios Fraccionales) en el modelo de Cox, como una alternativa en situaciones en que los efectos de las covariables no presentan una relación lineal en el logaritmo de la razón de riesgos.

1.2. Objetivos del trabajo

El objetivo de este trabajo final de máster es analizar el efecto de las covariables en la supervivencia de un grupo de pacientes con linfoma no Hodgkin, utilizando métodos flexibles como los P-splines y polinomio fraccional para aproximar el efecto de las covariables en el contexto del modelo de Cox, cuando el supuesto de los riesgos proporcionales no se verifica y el efecto de las covariables no presenta una estructura de relación lineal.

Los objetivos del análisis son dos:

-Determinar los factores pronósticos para la supervivencia global en pacientes con LNH. Muchas de las variables clínicas son factores pronósticos para la supervivencia de los pacientes con LNH, los cuales han sido tratados en la mayoría de los estudios como variables continuas categorizadas.

-Determinar la forma funcional del efecto de las covariables en la función de riesgo, 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, que comprenden los aspectos clínicos del linfoma no Hodgkin (LNH), conceptos básicos relacionados con el análisis de supervivencia, modelo de regresión de Cox con métodos flexibles, factores pronósticos en LNH, discusión y conclusiones.

En el capítulo 2 se presenta la descripción de los aspectos clínicos de los LNH y los datos que son objeto de análisis. En el capítulo 3 se presentan los conceptos básicos en el análisis de supervivencia. En el capítulo 4 se describe la base teórica del modelo de regresión de Cox utilizando los métodos flexibles. En la sección 4.1 se describe brevemente el modelo de Cox clásico, en la sección 4.2 se describe el modelo de Cox con método P-splines, en la sección 4.3 se describe el modelo de Cox con método de polinomio fraccional y en la sección 4.4 se describe los procedimientos de diagnóstico del modelo de Cox. En el capítulo 5 se describen los factores pronósticos en los LNH mediante el modelo de Cox utilizando P-splines y polinomio fraccional. En el capítulo 6 se da una breve discusión y las conclusiones de este trabajo.

Capítulo 2
LNH: Aspectos clínicos y metodológicos

En este capítulo se presenta los aspectos clínicos del Linfoma No Hodgkin, de manera que, nos permita nococer 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, principalmente de las variables explicativas, que son objeto de análisis para determinar los factores pronósticos, mediante métodos flexibles que son descritos en el siguiente capítulo.

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).

El linfoma 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 y 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)).

El tratamiento de los pacientes con LNH, depende del grado de agresividad (indolente y agresivo) y el estadio clínico de la enfermedad o el índice pronóstico internacional (IPI) que clasifica a los pacientes en cuatro grupos de riesgo (bajo riesgo, intermedio bajo, intermedio alto y alto riesgo) teniendo en cuenta la edad, estado funcional, estadio clínico, nivel deshidrogenasa láctica (DHL) 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. La quimioterapia a base del 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 sus 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)).

El pronóstico de los pacientes con LNH depende de múltiples factores, siendo los más relevantes: la edad, estado de performance (escala Karnofsky o ECOG), estadio clínico, DHL, β2-microglobulinas ($\beta 2M$), tipo histológico, tipo celular (linaje B o T) y grado de agresividad.

Según algunas series publicadas, los pacientes mayores de 60 años de edad, enfermedad ganglionar, escala ECOG 2-4, estadio clínico III-IV, síntomas B (fiebre, sudoración nocturna y baja de peso sin causa alguna), DHL elevada ($>240 U/L$) y $\beta 2M$ elevada (>3.5), linfoma agresivo, linfoma de células T, y en algunas series pacientes con hemoglobina baja ($<12 g/dl$), leucocitos elevados ($> 10 mil$) 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 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 real, debido a la aparición de nuevos FP (hemoglobina, leucocitos, linfocitos y marcadores tumorales), la existencia de un número relativamente importante de pacientes que presentan recurrencia o que fallecen a consecuencia de la enfermedad 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 del mismo 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, estado funcional, 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 ó 70 años. La capacidad funcional del paciente, según la escala ECOG (Zubrod), 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 del paciente que puedan influir en la morbimortalidad y tolerancia al tratamiento se consideran como factores pronósticos (ejemplo, las 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 (células B o T), las alteraciones citogenéticas, actividad proliferativa, extensión de la enfermedad y otras variables con significado pronóstico. El subtipo histológico, el patrón de infiltración (folicular o difuso) y aspecto citológico de las células así como su diferenciación que dividen a los LNH en grupos diferenciados se consideran como factores pronósticos; sin embargo, el pronóstico de las diferentes entidades no son sustancialmente diferentes. El estudio del inmunofenotipo, ha permitido establecer el significado pronóstico, al demostrarse 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, mientras la ausencia se ha visto asociada a una mayor supervivencia.
  La extensión de la enfermedad, definida como la cantidad del tumor al momento del diagnóstico, reviste una 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 ("masa Bulky", aquella masa cuyo tamaño del diámetro mayor es superior a $10cm$), carga tumoral (número de regiones ganglionares extensas (bulky) 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 β2-microglobulina elevada.

2.4. Descripción de los datos

La base de datos objeto de análisis, corresponde a los datos de 2160 pacientes mayores o iguales a 14 años de edad con diagnóstico de LNH, que fueron diagnosticados 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 de los LNH.

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

2.4.1. Descripción de los datos.

Los datos recopilados al diagnóstico de las variables relacionadas a las características del paciente y del tumor (características clínicas) fueron:

- Edad: en años.
- Género: Femeno o masculino.
- Zubrod: Estado funcional del paciente, según la escala ECOG.
- Primario: Localización ganglionar o extraganglionar.
- Tumor: Diámetro mayor del tumor.
- Numero de ganglios afectados.
- Numero de sitios extraganglionares.
- Estadio clínico: Extensión de la enfermedad, según la clasificación Ann Arbor.
- Sitios de metástasis: Extensión del tumor a otras regiones o órganos.
- Síntomas: Fiebre, sudoración nocturna o baja de peso sin causa alguna.
- Tipo de LNH: Subtipo histológico, según la clasificación disponible.
- VIH/SIDA: Infección por VIH o SIDA.
- Hemoglobina: en g/dl.
- Leucocitos: Número de leucocitos $/mm^3$.
- Linfocitos: Linfocitos en porcentaje.
- Deshidrogenasa láctica: en UI/L.
- β 2-microglobulina: en mg/L

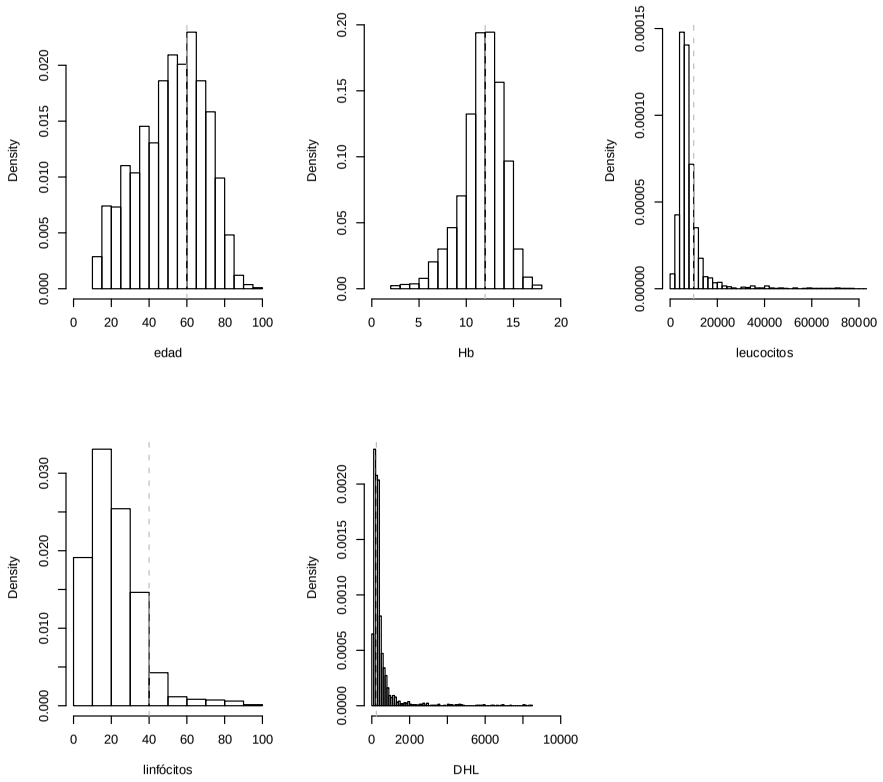
De los cuales, no se incluye en el análisis la información de las siguientes variables: tamaño del tumor, número de ganglios y el sitio de metástasis, debido a que estas variables ya están reflejadas en el estadio clínico. Así mismo, no se incluye el subtipo histológico y el genotipo (células T o B) debido a que los pacientes fueron diagnosticados con tres criterios de clasificación histopatológica diferentes (Rappaport y Kiel, formulación de trabajo (WF) y la clasificación REAL). Y la $\beta 2M$, debido a que este dato había sido solicitado en muy pocos pacientes.

En la Tabla 1 se muestra el número de casos por cada una de las categorías de las variables evaluables. El 36.9 % eran mayores de 60 años de edad, 50.9 % eran de sexo masculino, 27.0 % presentaban zubrod entre 2-4, 66.7 % tenían enfermedad ganglionar, 49.2 % presentaban enfermedad en estadio clínico avanzado (EC IIIIV), 38.1 % habían presentado síntomas B, 48.1 % habían presentado hemoglobina bajo (Hb $<12 g/dl$), 17.7 % leucocitos elevados (leucocitos >10mil), 7.7 % linfocitos elevados (linfocitos > 40 %) y 60.0 % niveles de DHL elevados ($>240\ UI/L$).

Tabla 1. Variables y número de casos por cada categoría.

Variables	Mediana/rango	Categoría	Casos	Porcentaje (%)
Edad (años)	54.0 / 14 - 96	$\leq$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	-	Extragang	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	$\geq$12	1122	51.9
		<12	1038	48.1
Leucocitos ($10^3/mm^3$)	6.7 / 0.560 - 163	$\leq 10^3$	1777	82.3
		$> 10^3$	383	17.7
Linfocitos (%)	20 / 1 - 94	$\leq$40	1993	82.3
		>40	167	7.7
Deshidrogenasa láctica (UI/L)	298 / 24 - 8440	$\leq$240	863	40.0
		>240	1297	60.0

Por otro lado, en la Gráfica 1 se muestra la distribución de las covariables continuas y se observa que todas ellas presentan asimetría. La asimetría es más pronunciada para las variables leucocitos y la deshidrogenasa láctica (DHL); las cuales, son posibles debido a que los valores de los leucocitos pueden varíar entre < 1000 y 100mil $/mm^3$ y de los DHL entre <240 y $10 mil\ UI/L$.



Gráfica 1. Distribución de las variables continuas de los datos de pacientes con LNH.

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 el método 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 o tiempo 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, progesión, 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.

Para propósitos del presente trabajo nos enfocaremos en datos con censura o censurados por la derecha y de tipo aleatorio. La censura 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 siguimiento, finalización del estudio), de modo que los tiempos de censuramiento son también aleatorias.

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

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

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

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

De la expresión (1) y (2) la función de riesgo es, $\lambda(t) = f(t)/S(t)$. Por otro lado, 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

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

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.

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, debido a cualquiera de las tres causas descritas anteriormente (perdida de acompañamiento, retiro del estudio por eventos ajenos al estudio y término del estudio), 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, mediante el método de máxima verosimilitud para los parámetros del modelo o la función de supervivencia bajo censuramiento por la derecha y de manera aleatoria.

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, δ) 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 δ es la variable indicadora de censura.

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

$$L = \prod_{i=1}^{n} [f(t_i)(1 - G_i)]^{\delta_i} [g_i S(t_i)]^{1-\delta_i}. \tag{3}$$

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:

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

En consecuencia, reemplazando las funciones respectivas en las expresiones (3) y (4), se pueden obtener los estimadores de los parámetros del modelo de distribución supuesto para variable aleatoria T y 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 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.

En esta sección se describe brevemente los términos básicos del proceso de conteo.

3.4.1. Procesos de conteo.

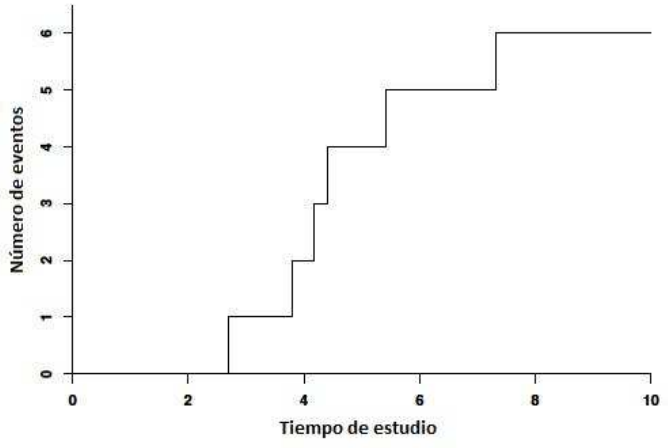
Los tiempos de supervivencia pueden ser representados a través de ciertos procesos estocásticos. Los datos en sí pueden ser descritos como un proceso de conteo, el cual es simplemente una función aleatoria del tiempo t, $N(t)$. Esta función es cero en el tiempo inicial y constante en el tiempo, excepto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

Considerando para un solo tipo de evento. Para un tiempo t dado, sea $N(t)$ el número de eventos que han ocurrido hasta el tiempo. 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 2.

En la Gráfica 2 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.



GRÁFICA 2. Ilustración de un proceso 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, la función de intensidad al tiempo t es definido 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:

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

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 el 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.

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, δ_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 a 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

$$\begin{aligned} &P[t \leq T_i \leq t + dt, \delta_i = 1 | \mathfrak{F}_{t-}] \\ &= \begin{cases} P(t \leq T_i \leq t + dt, C_i > t + dt_i | T_i \geq t, C_i \geq t) = h(t)dt & \text{si} \quad T_i \geq t \\ 0 & \text{si} \quad T_i < t \end{cases} \end{aligned} \tag{5}$$

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, $dN(t) = N[(t + dt)^-] - N(t^-)$ (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 (5),

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

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,

$$E(dM(t)|\mathfrak{F}_{t^-}) = E[dN(t) - d\Lambda(t)|\mathfrak{F}_{t^-}] = 0 \tag{6}$$

La fórmula (6) 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, $\lambda(t) = (\lambda_1(t), \ldots, \lambda_n(t))$ un proceso de intensidad n-dimensional, $M(t) = (M_1(t), \ldots, M_n(t))$ 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 definidas 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, es decir la función no asume una forma funcional o 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 $f(\cdot)$ 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 con 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, pero la selección de los nodos se hace mediante algoritmos bastante complicados. Como una alternativa a estas complicaciones de ambos métodos sea 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 p 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 p
- La función $s(x)$ tiene $(p-1)$ 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 p 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 $p + k + 1$.

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

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

$$S(x) = \sum_{j=0}^{p} \beta_j X^j + \sum_{j=1}^{k} \lambda_j (x - \xi_j)_+^m \tag{7}$$

donde la notación

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

De la expresión (7), el caso más simple es un spline lineal. 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:

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

En general, los tipos de splines dependen del tipo de las funciones bases que se utilizan.

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 splines penalizados.

3.5.3. 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 como $\tau_1 = \ldots = \tau_M = t_0, t_{k+1} = \tau_{k+M+1} = \ldots = \tau_{k+2M}$.

Se renombra los nodos originales como $\tau_{M+j} = t_j$, $j = 1, \ldots, k$.

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

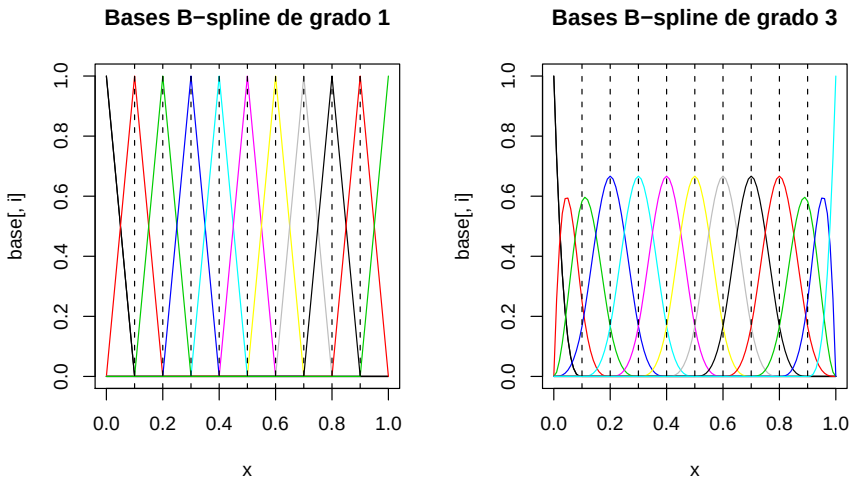
Para $m = 2, \ldots, M$, los B-splines de orden m se definen como

$$B_{j,m} = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_{j,m-1} + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1,m-1}. \tag{8}$$

En la Gráfica 3 se muestra las 13 funciones que forman las bases B-splines de grado 1 y 3 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 dos nodos internos.

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 $[t_j, t_{j+m}]$
- $\sum B_j(x) = 1$



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

3.5.4. P-splines.

Eilers y Marx (1996) introducen el término P-splines, llamado splines penalizado (Ruppert y Carroll, 2000) o pseudo-splines (Hastie, 1996), son una extensión de 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:

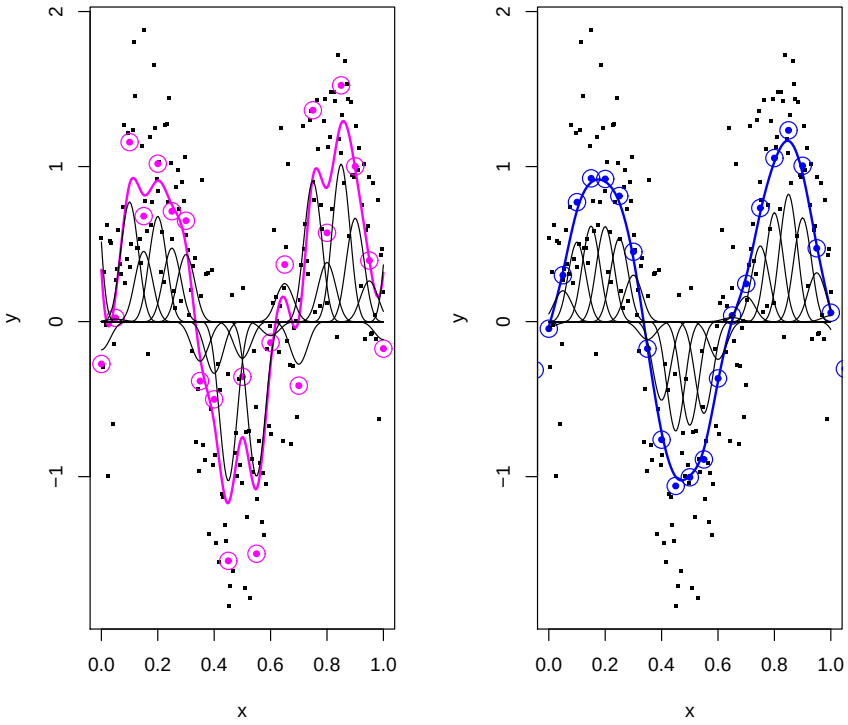
$$f(x) = \sum_{m=1}^{M} \alpha_m B_m(x), \tag{9}$$

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),

$$(1/2)\lambda \int [f''(x)]^2 dx. \tag{10}$$

Supongamos que tenemos una base B-splines construida con k nodos y utilizamos mínimos cuadrados penalizados para ajustar el modelo. La Gráfica 4 muestra ajuste de una curva mediante B-splines sin y con penalización, las funciones que forman la base multiplicadas por los coeficientes, así como los coeficientes (en círculo). El efecto de la penalización que 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 y con penalización.

Capítulo 4
Modelo de regresión de Cox con métodos flexibles

En este capítulo se presenta una breve revisión de literatura y se describe la base teórica del modelo de regresión de Cox utilizando los métodos flexibles mediante los splines penalizados y el polinomio fraccional. Así mismo, se describe los procedimientos de diagnóstico del modelo de Cox.

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 semiparamétricos 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 suavización (smoothing spline) y polinomios fraccionales (fractional polynomials o FP). En este trabajo se utilizan estos dos métodos de modelamiento 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 (penalized splines o 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. En cambio, los FP (Royston y Sauerbrei, 2008) aproximan la función desconocida por una suma de trasformaciones potencia de las covariables, que son más flexibles que los polinomios ordinarios, como tal admiten potencias negativas y no enteras.

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 (penalized splines, P-splines), un enfoque diferente que puede ser visto como un compromiso entre suavizamiento spline y regresión spline.

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 con métodos flexibles para determinar los factores pronósticos en la supervivencia global de los pacientes con LNH y se estima la forma funcional del efecto de las covariables en la razón de riesgo (hazard ratio). Previamente se hace una breve descripción del modelo de Cox (modelo de Cox clásico), luego se describe el modelo de Cox con P-spline y modelo de Cox con polinomio fraccional, y finalmente se realizan las comparaciones entre los dos métodos (P-splines y FPs), en 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,δ,X), donde, T es el tiempo de supervivencia observada, δ 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

$$\lambda(t/X) = \lambda_0(t)\exp\{\mathbf{X}'\beta\}, \tag{11}$$

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 (11):

- 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 ($forma\ funcional\ lineal$).

En el modelo de la forma (11), el problema de modelamiento de los datos de supervivencia se reduce a la estimación de los parámetros β 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, δ_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 (11) la estimación de los parámetros β 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 β, 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

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(X'_{(j)}\beta)}{\sum\limits_{l \in R_j} \exp(X'_l\beta)} \tag{12}$$

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 (11) para cada sujeto i y definida la probabilidad de observar una falla en un individuo i de la forma

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

el logaritmo de la función de verosimilitud parcial es dado por

$$\ell(\beta) = ln[L(\beta)] = \sum_{l=1}^{n} \delta_{(i)} \left\{ X'_{(l)}\beta - ln \sum_{l \in R(t_i)} \exp(X'_l\beta) \right\} \tag{13}$$

Cuando los datos contienen tiempos observados empatados, la verosimilitud parcial (12) tiene que ser 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

$$\prod_{j=1}^{r} \frac{\exp(s_j'\beta)}{\left[\sum\limits_{l\in R_j} \exp(s_l'\beta)\right]^{d_j}} \tag{14}$$

Una aproximación sugerida por Efron (1977) es

$$\prod_{j=1}^{r} \frac{\exp(S_j'\beta)}{\prod_{k=1}^{d_j}\left[\sum_{l\in R_j}\exp(S_l'\beta) - (k-1)d_k^{-1}\sum_{l\in D_j}\exp(X_l'\beta)\right]} \tag{15}$$

Otra aproximación sugerida por Cox (1972) es

$$\prod_{j=1}^{r} \frac{\exp(S_{(j)}'\beta)}{\sum\limits_{l\in R(t_{(j)};d_j)}\exp(S_l'\beta)} \tag{16}$$

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 expresión (12), (13), (14), (15) y (16), 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, β_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 (11) 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 χ^2 con p grados de libertad.

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 $I(\beta)$, $\chi^2_{Sc} = U(\beta_0)' I^{-1}(\beta_0)U(\beta_0) \sim \chi^2_p$

Por otro lado, debido a que los métodos de diagnóstico del modelo están desarrollados en el contexto de proceso de conteo y teoría de martingalas, 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, δ_i, X_i), donde T_i es el tiempo de supervivencia censurado por la derecha, δ_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

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

donde $Y(t)$ es una 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 β 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 (17).

Los parámetros β 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 β.

Si β es fijado, el estimador Nelson-Aalen de $\Lambda_0(t)$ es estimado como $\hat{\Lambda_0}(t,\beta) = \int_0^t \frac{1}{S_0(s,\hat{\beta})} dN(s)$, $N(t) = \sum_t N_i(t)$. 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 β. Sea $I(\beta)$ el negativo de la primera derivada de la función score con respecto a β y sea β_0 que denota el verdadero valor de β.

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.

4.3. Modelo de regresión de Cox con P-splines

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

$$h(t|X) = \lambda_0(t) \exp(g(X)), \tag{18}$$

donde g es una función no especificada.

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 (18) 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.

En este sentido el modelo de regresión aditivo (Stone, 1985) facilita una estructura que permite diferentes funciones para cada covariable. Por lo tanto, en el modelo de Cox, el logaritmo de la razón de riesgo tiene p componentes, cada uno representado por una función arbitraria:

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\{g_1(X_1) + \ldots + g_p(X_p)\}. \tag{19}$$

En el modelo (19) las p funciones desconocidas pueden ser aproximadas mediante splines,

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\{s_1(X_1) + \ldots + s_p(X_p)\}. \tag{20}$$

Los splines son capaces de aproximar las funciones conocidas, por tanto, se espera lo mismo para las funciones componentes en (20).

Considerando las componentes de covariables binarias y aquellas que satisfacen una forma funcional lineal y las componentes que no satisfacen la estructura lineal, el modelo (20) puede ser expresado en una forma más general como

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\left\{\sum_{j=1}^{p}\beta_j X_j \; + \sum_{j=p+1}^{p+q} g_j(Z_j)\right\}. \tag{21}$$

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

En el modelo (21) las q covariables que no satisfacen la forma funcional lineal pueden ser aproximadas mediante regresión splines o suavizamiento splines.

Aquí se describe brevemente los procedimientos de aproximación de las funciones g mediante regresión splines y utilizando suavizamiento splines con penalización, aunque la aplicación se realiza sólo para P-splines.

Sleeper y Harrington (1990) utilizan regresión splines para aproximar las funciones desconocidas, g, en el modelo (21). Si bien la transformación g de una covariable puede ser aproximada por un polinomio, esta es sensible a outliers; sin embargo, los outliers tienen un efecto más local cuando se utiliza aproximación mediante polinomio por trozos (piecewise polinomial). Un spline es un polinomio por trozos con la condición de continuidad en la función y sus derivadas en los puntos donde los trozos se unen.

Sea $X = x$ una covariable en el modelo de Cox (20), que es trasformada en un vector d-dimensional de bases B-splines, $[B_1(x), ..., B_d(x)]$, de manera que $s(x) = \sum_{l=1}^{d}\alpha_l B_l(x)$.

Entonces el modelo spline se puede escribir como

$$\lambda(t|\mathbf{X} = \mathbf{x}) = \lambda_0(t)\exp\left\{\sum_{l=1}^{d}\alpha_l B_l(x)\right\} \tag{22}$$

donde $\alpha = (\alpha_1, \ldots, \alpha_d)$ es el vector de coeficientes de las bases B-splines que puede ser estimado de manera similar a d coeficientes de regresión en el modelo de Cox.

Aunque el modelo (22) es sobreparametrizado, esta sobreparametrización podría resolverse quitando una de las bases B-splines del modelo (Sleeper y Harrington, 1990). Por lo tanto, en el modelo (22) el logaritmo de la razón de riesgos se puede estimar mediante $\hat{s}(x) = \sum_{l=1}^{d} \hat{\alpha}_l B_l(x)$.

Considerando el modelo de la forma más general, parcialmente lineal (21),

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp\left\{ \sum_{j=1}^{p} \beta_j X_j \; + \sum_{j=p+1}^{p+q} s_j(Z_j) \right\}. \tag{23}$$

donde las p covariables son categóricas o por otras razones no son trasformados en bases splines y las siguientes q covariables son transformadas en vectores de bases B-splines de longitudes $d_1, \ldots, d_q$,

$[B_{11}(Z_{p+1}), \ldots, B_{1d_1}(Z_{p+1})], \ldots, [B_{q1}(Z_{p+q}), \ldots, B_{qd_q}(Z_{p+d_q})]$, con $d = d_1 + \ldots + d_q$ número total de coeficientes de las bases B-splines.

Finalmente el modelo de Cox parcialmente lineal (23) tiene $p + d$ coeficientes de regresión $\alpha = [\alpha_1, \ldots, \alpha_p, \alpha_{p+1}, \ldots, \alpha_{p+d_q}]$ que pueden ser estimados utilizando los procedimientos tradicionales como el método de máxima verosimilitud parcial para el modelo de Cox.

En consecuencia, los parámetros se pueden estimar como en el modelo de regresión clásica y realizarse las pruebas de hipótesis $H_0 : \alpha = 0$, mediante los estadísticos de prueba tipo Wald, el de razón de verosimilitud o score.

Sin embargo, este método es más sensible al número y la localización de los nodos que los splines penalizados, y por tanto, son mas inestables.

En el package R, la función coxph permite realizar el ajuste del modelo de Cox con regresión spline, por ejemplo B-splines natural (ns, df=4).

4.3.1. Modelo de Cox con P-splines.

Gray (1992) utiliza splines penalizados para aproximar las funciones desconocidas, g, en el modelo (21). 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.

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

Sea el modelo de la forma más general (21), que incluye términos lineales y no lineales (modelo parcialmente lineal), que puede ser expresado de la forma (por simplicidad)

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\left\{\sum_{j=1}^{p}\beta_j X_j \ + \sum_{k=1}^{q} s_k(z_k)\right\}. \tag{24}$$

donde cada una de las q componentes no lineales $s_k(z_k)$, es aproximado por una combinación lineal de las funciones bases B-spline cúbico (deBoor, 1978),

$$s_k(z_k) = \theta_{k,0}z_k + \sum_{m=1}^{M+2}\theta_{k,m}B_{k,m}(z_k) \tag{25}$$

Una función base B-spline cúbico con M nodos tiene $M+4$ funciones, debido a que el espacio de estas funciones incluye el término constante y lineal. En la expresión (24) 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 usada es dada por

$$\frac{1}{2}\lambda_k \int [s_k^{''}(z_k)]^2 dz, \tag{26}$$

donde λ_k es un parámetro de suavizamiento que controla el grado de suavidad.

En este sentido, 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 (24) con términos no lineales parametrizados por (25) y utilizando la penalización (26) 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.

4.3.2. Estimación de los parámetros.

Sean los parámetros de los términos no lineales denotados como $\theta_k = (\theta_{k,1}, \ldots, \theta_{k,M+2})$ y $\vartheta_k = (\theta_{k,0}, \theta_k)$. La expresión (26) es cero cuando s_k es lineal y aumenta a medida que s_k se hace menos suave. En la función de penalización solo aparecen los parámetros θ_k y (26) es una función cuadrática en los parámetros, por lo que se pude escribir como

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

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 el resto de la matriz.

Denotando al conjunto de parámetros β y ϑ por $\eta=(\beta,\vartheta)$, y sea $\ell(\eta)$ el logaritmo de la verosimilitud parcial (Cox, 1972) para el modelo (24) con s_k parametrizado por (25), entonces el logaritmo de la verosimilitud parcial penalizado ($\ell_p(\eta)$) es,

$$\ell_p(\eta)=\ell_p(\eta)-\frac{1}{2}\sum_{k=1}^{q}\lambda_k\theta_k' P_k\theta_k, \tag{27}$$

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

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 (27), 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.

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 y $I(\eta,\lambda)=\left(\frac{d^2\ell_p(\eta)}{d^2\eta}\right)$ la matriz de información. Bajo la normalidad asintótica de los estimadores de máxima verosimilitud parcial, los estadísticos de prueba pueden ser formadas exactamente como en el análisis de verosimilitud ordinario utilizando estas cantidades para realizar las distintas pruebas de hipótesis sobre los parámetros del modelo aditivo (24).

4.3.3. Prueba de hipótesis e inferencia.

En el modelo de la forma (24) existen tres hipótesis de prueba, aunque dos de ellas referidas a los términos no lineales son de interés particular:

- Hipótesis sobre el efecto global, $H_0:\eta=0$ vs. $H_1:\eta\neq 0$
- Hipótesis que la k-ésima covariable no tiene efecto, $H_{01_k}:\vartheta_k=0$ vs. $H_{01_k}:\vartheta_k\neq 0$
- Hipótesis que la k-ésima covariable tiene efecto lineal $H_{02_k}:\theta_k=0$ vs. $H_{02_k}:\theta_k\neq 0$

Para una muestra de tamaño n suficientemente grande, Sea $(\hat\beta,\hat\vartheta)$ los valores de los parámetros que maximizan el logaritmo de la verosimilitud $\ell_p(\eta)$. Bajo la distribución asintótica de los estimadores de máxima verosimilitud, se pueden construir las pruebas de hipótesis sobre los parámetros del modelo (24).

Considerando primero la hipótesis $H_0 : \vartheta = 0$. Sea $\hat{\beta}_0$ el estimador de la máxima verosimilitud parcial para β 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 ϑ. 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 score penalizado: $Q_{Sc} = U'_\vartheta(\hat{\beta}_0, 0)(I_{\vartheta\vartheta/\beta} + \lambda P^*)U_\vartheta(\hat{\beta}_0, 0)$, donde $I_{\vartheta\vartheta/\beta} = I_{\vartheta\vartheta} - I_{\vartheta\beta}I^{-1}_{\beta\beta}I_{\beta\vartheta}$.
- Prueba de razón de verosimilitud penalizado: $Q_{LR} = 2[\ell_p(\hat{\beta}, \hat{\vartheta}) - \ell_p(\hat{\beta}_0, 0)]$.
- Estadístico de prueba tipo Wald: $Q_W = \hat{\vartheta}'(I_{\vartheta\vartheta/\beta} - \lambda P^*)\hat{\vartheta}$.

La construcción de los estadísticos de prueba para la hipótesis de que la k-ésima covariable presenta un efecto lineal, $H_{0_k} : \theta_k = 0$, se puede hacer exactamente lo mismo, pero con θ_0 y β en lugar de ϑ.

En el programa R, la función coxph y pspline de la librería package survival permiten realizar el ajuste del modelo de Cox, con parte de los componentes no lineales aproximado mediante splines penalizados. Los estadísticos de prueba disponibles son: la prueba de razón de verosimilitud y la prueba tipo-Wald.

4.4. Modelo de regresión de Cox con polinomio fraccional

Royston y Altman (1994) introducen el modelo de regresión usando polinomios fraccionales para covariables continuas. El polinomio fraccional fue introducido como una extensión del método de polinomio para modelar con predictores continuos. Detalles técnicos de cómo es el modelo de polinomio fraccional para un predictor y cómo las pruebas de significación se determinan es detallado en Royston y Altman (1994). Una introducción del polinomio fraccional en el contexto de estimación de factores pronósticos en pacientes con cáncer de mama es dado por Sauerbrei y Royston (1999).

Sauerbrei y Royston (1999) proponen la aproximación con MFP (Multivariable Fractional Polinomials) que fue desarrollado posteriormente para incluir en el modelo predictores de formas diferentes, donde al menos uno es continuo. El método fue aplicado en el modelamiento de pronóstico y diagnóstico en cáncer de mama (Sauerbrei y Royston, 1999).

Sea el modelo de Cox de la forma general (21) de manera que cada componente no lineal g, puede ser aproximado por una suma de transformaciones de potencias.

Los FP's aproximan cada función desconocida g por una combinación lineal de M polinomios x^{p_j}, $j = 1, \ldots, M$.

En el polinomio ordinario las potencias p_j son restringidas para valores enteros positivos, mientras en modelamiento con FP's se trabaja con valores positivos, no-positivos y valores fraccionales para p_j. Un típico conjunto de potencias admisibles es dado por $p_j \in (-2, -1, -0{,}5, 0{,}5, 1, 2, 3)$, donde x^0 denota $ln(x)$.

Más formalmente, un FP de grado M es definido como

$$FP_M(x) = \sum_{j=1}^{M} \beta_j h_j(x),$$

donde $\beta_1, \ldots, \beta_M$ son los coeficientes de regresión y h_j es recursivamente definido como

$$h_0(x) = 1,$$

y

$$h_j(x) = \begin{cases} x^{p_j}, & \text{si } p_j = p_{j-1} \\ h_{j-1}(x) ln(x), & \text{si } p_j \neq p_{j-1} \end{cases}$$

Para $M = 2$ y $p_j \neq p_{j-1}$: $FP_2(x) = \beta_1 x^{p_1} + \beta_2 x^{p_2}$.

Para $M = 2$ y $p_j = p_{j-1}$: $FP_2(x) = \beta_1 x^{p_1} + \beta_2 x^{p_1} ln(x)$.

Programas para ajustar el modelo aditivo basado en FPs es evaluable en la plataforma de computación estadística STATA (función mfp), SAS (macro mfp8) y R (función fp del paquete mfp). La implementación en R se restringe para FP's de grado 2 (Sauerbrei et al., 2006).

4.5. 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 (11), las restricciones naturales suponen verificar los siguientes aspectos:

- El logaritmo de la razón de riesgo no depende del tiempo, $ln(\lambda(t, x)/\lambda_0(t)) = \beta_1 X_1 + \ldots + \beta_p X_p$ (riesgo constante).
- El riesgo de un individuo es proporcional al riesgo de otro individuo (riesgo proporcional).
- El logaritmo de la razón de riesgo y las covariables se relacionan linealmente (forma funcional lineal).

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.

4.5.1. Verificación del supuesto de riesgos proporcionales.

4.5.1.1. 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),

$$\lambda(t) = Y(t)\lambda_0(t)\exp(X'(t)\beta(t)) \tag{28}$$

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 el tiempo sería una línea horizontal.

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 $(k = 1, \ldots, d)$,

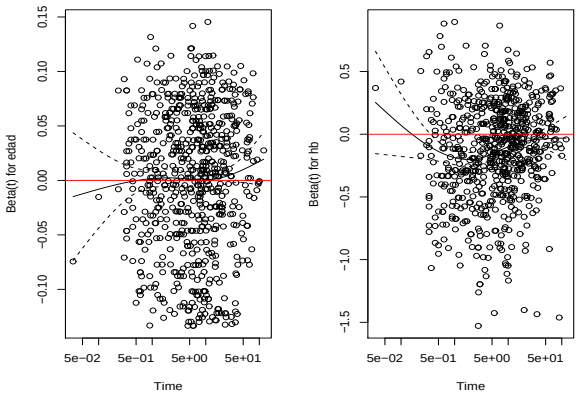
$$\widetilde{r}_k = \widetilde{X}_k - E(\widetilde{X}_k/R_k),$$

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, β_{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



GRÁFICA 5. Residuos Schoenfeld escalado vs. $g(t) = log(t)$, para la edad y Hb.

riesgos proporcionales, los residuos se distribuyen alrededor de la línea horizontal para un coeficiente β_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 5, 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 ± 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.

4.5.1.2. 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 (28) pueden ser escritos como,

$$\beta_j(t) = \beta_j + \theta_j(g_j - \bar{g}_j), j = 1, \ldots, p \tag{29}$$

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 (29) 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

$$T_j = \frac{\left(\sum (g_k - \bar{g}) r s_{kj}\right)^2}{d I_{jj} \sum (g_k - \bar{g})^2},$$

  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 χ^2_1.

Por otro lado, sean los coeficiente de regresión tiempo-dependiente del modelo de Cox extendido (28) que puede ser escrito como

$$\beta_p(t) = \beta_p + \theta_p g_p(t), \tag{30}$$

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

Cuando las $g's$ son funciones conocidas, entonces el modelo con coeficientes (30) es aún el modelo de Cox. Por tanto, el estimador de los parámetros se puede obtener maximizando la función de verosimilitud partial y las pruebas de hipótesis se pueden realizar 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 β y la segunda componente respecto a θ. 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 (30), se usa 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 χ^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 5), 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).

4.5.2. Verificación de la forma funcional lineal.

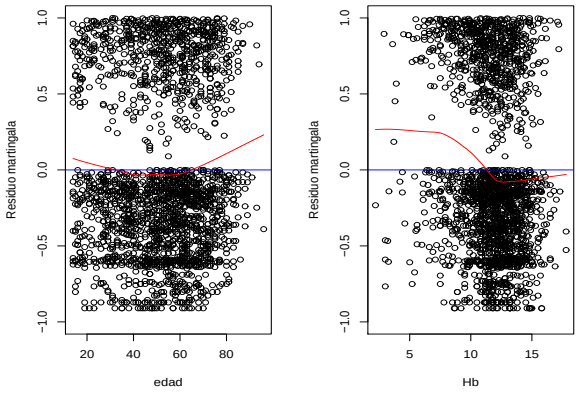
En el modelo de Cox, uno de los residuos muy usuales para verificar la forma funcional del efecto de las covariables en la función de riesgo son los residuos basados en martingalas. Barlow y Prentice (1988) provee el marco básico de los residuos martingala y el posterior trabajo de Therneau, Grambsch 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 (17) 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 δ_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 6 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)



GRÁFICA 6. Residuos martingala vs. edad y Hb.

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).

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, 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.

Capítulo 5
Factores pronósticos en LNH

5.1. Descripción de los datos

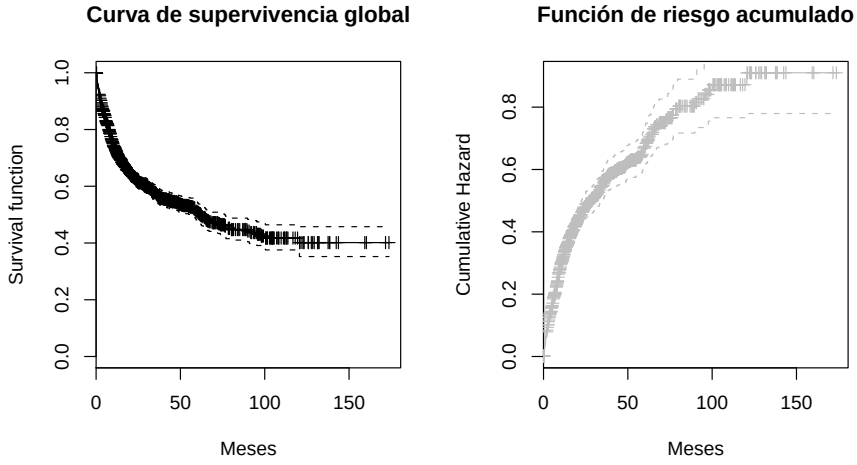
En este trabajo se analizan los datos de 2160 pacientes mayores o iguales a 14 años de edad con diagnóstico de linfoma no Hodgkin (LNH) que fueron diagnósticados y tratados en el Instituto Nacional de Enfermedades Neoplásicas (INEN), Lima-Perú, entre 1990 y 2002.

El tratamiento que habían recibido los pacientes según la práctica clínica habitual (según el protocolo de tratamiento) 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.

Las siguientes características clínicas (covariables) de los pacientes, documentados al diagnóstico, fueron incluidos en el análisis: la edad (en años), género (femenino, masculino), estado funcional (zubrod: 0, 1, 2, 3, 4), foco primario (primario: enfermedad ganglionar, extraganglionar), estadio clínico (EC: I, II, III, IV), síntomas B (fiebre, sudoración nocturna o baja de peso, sin causa alguna), hemoglobina (Hb: en g/dl), leucocitos (leuco: en mil/mm^3), linfocitos (linf: en %), y deshidrogenasa láctica (DHL: en UI/L). No se incluye la β-2 microglobulina (β2M: en mg/L), debido a que la mayoría de los pacientes no tenían información al respecto.

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. Los pacientes fallecidos se consideraron como eventos (no censurados) y los restantes como censurados.

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 fue de 61.8 meses (IC95: 49.9 - 73.7) y la tasa de supervivencia a 5 y 10 años de 51.2 % y 41.7 % respectivamente. En la Gráfica 7 se muestra la curva de supervivencia estimada mediante el método de Kaplan-Meier y la función de riesgo acumulado.



GRÁFICA 7. Curva de supervivencia y riesgo acumulado de los pacientes con LNH.

5.2. Aplicando el modelo de Cox clásico

En la Tabla 2 se muestran los resultados de la aplicación del modelo de Cox, 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, ln(leucocitos), linfocitos y ln(DHL).

TABLA 2. Resultados del 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)
Hb	-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	351.00				
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 2) se observa que todas las covariables, a excepción del foco primario (p = 0.580), presentan un efecto significativo (p<0.05) en la supervivencia de los pacientes con LNH. La razón de riesgo de estas variables implican 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). Para el logaritmo de los leucocitos, el riesgo de mortalidad se incrementa en HR=1.24 por cada unidad que incrementa los leucocitos, así mismo, para el logaritmo del DHL el riesgo de mortalidad se incrementa en 1.3 por cada unidad que incrementa la DHL.

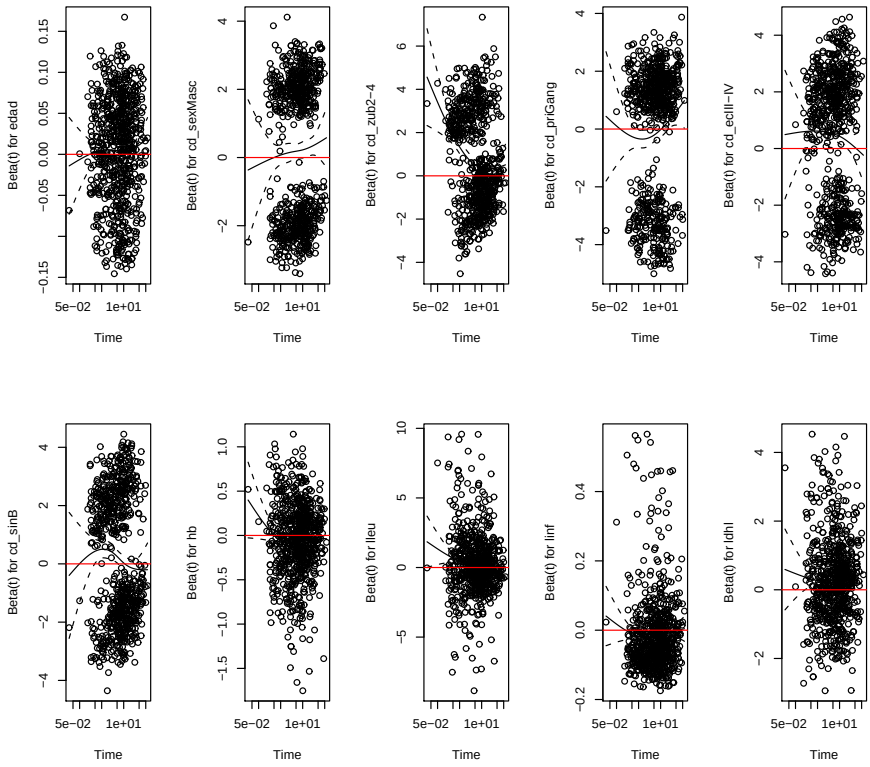
De acuerdo a los procedimientos definidos en la sección 4.5, se verifica el supuesto de riesgo proporcional basado en los residuos de Schoenfeld escalado y test de no proporcionalidad de Therneau y Grambsch y la forma funcional del efecto de las covariables en la función de riesgo basados en los residuos martingalas.

La Gráfica 8 muestra los residuos de Schoenfeld escalados para cada covariable en el modelo de Cox versus tiempo (ln(meses)). En los gráficos se observa que los residuos (líneas negras) no se aproximan a una línea horizontal (líneas rojas) en el punto 0 del eje Y que verifique el cumplimiento de los supuestos de riesgos proporcionales para las siguientes variables: zubrod (cd_zub2-4), primario (cd_priGang), estadio clínico (cd_ecIII-IV), síntomas (cd_sinB) y leucocitos (cd_lleu).

En la Tabla 3 se muestran los resultados del test de no proporcionalidad de Therneau y Grambsch. En los resultados se observa que el efecto de las covariables no es constante (p < 0.05). Por tanto, los resultados de las estimaciones bajo el modelo de Cox son discutibles, debido a que el modelo no cumple el supuesto de riesgos proporcionales (Test de no proporcionalidad global: p < 0.001).

TABLA 3. Prueba individual y global para la proporcionalidad, basado en los residuos de Schoenfeld escalados del modelo de Cox. La correlación de Pearson es entre los residuos y $g(t)$ para cada covariable.

Variables	correlación de Pearson (rho)	χ^2	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

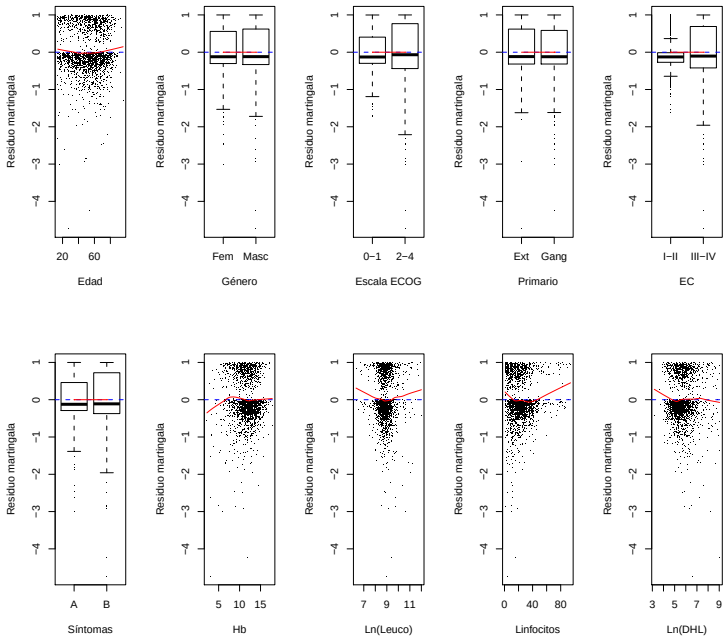


GRÁFICA 8. Residuos de Schoenfeld escalados vs. tiempo (ln(meses)) para el modelo de Cox clásico. En la primera fila de izquierda derecha se muestra para la edad, género, zubrod, primario y estadio clíncio. En la segunda fila de izquierda a derecha se muestra para síntomas, Hb, ln(leucocitos), linfocitos y ln(DHL)

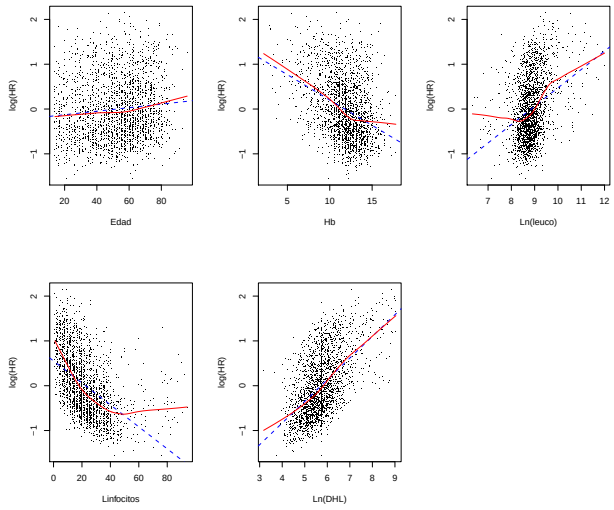
En la Gráfica 9 se muestran los residuos martingala para el modelo de Cox clásico. Los resultados indican que los residuos para cada covariable no son constantes. Las aproximaciones de los residuos muestran una tendencia no lineal para las variables continuas (líneas rojas); es decir, los efectos de las covariables continuas en el modelo no presentan un efecto lineal.

En la Gráfica 10 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 logaritmo de la razón de riesgo no presenta una relación lineal. Este resultado significa que estas variables no pueden entrar en el modelo en su escala original.

Si bien las trasformaciones de las variables pueden ser un recurso, en los análisis siguientes aproximamos la forma funcional del efecto de las covariables mediante métodos más flexibles como los splines y el polinomio fraccional.



GRÁFICA 9. Residuos martingala del modelo de Cox clásico.



GRÁFICA 10. Ln(HR) vs. las covariables: forma funcional de los efectos bajo el modelo de Cox clásico.

5.3. Aplicando el modelo de Cox con P-splines

En la Tabla 4 se muestra los resultados del modelo de Cox utilizando los splines penalizados (P-splines) 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 como el género, zubrod, primario, estadio clínico y síntomas.

Tabla 4. Resultados del modelo de Cox con P-spline.

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

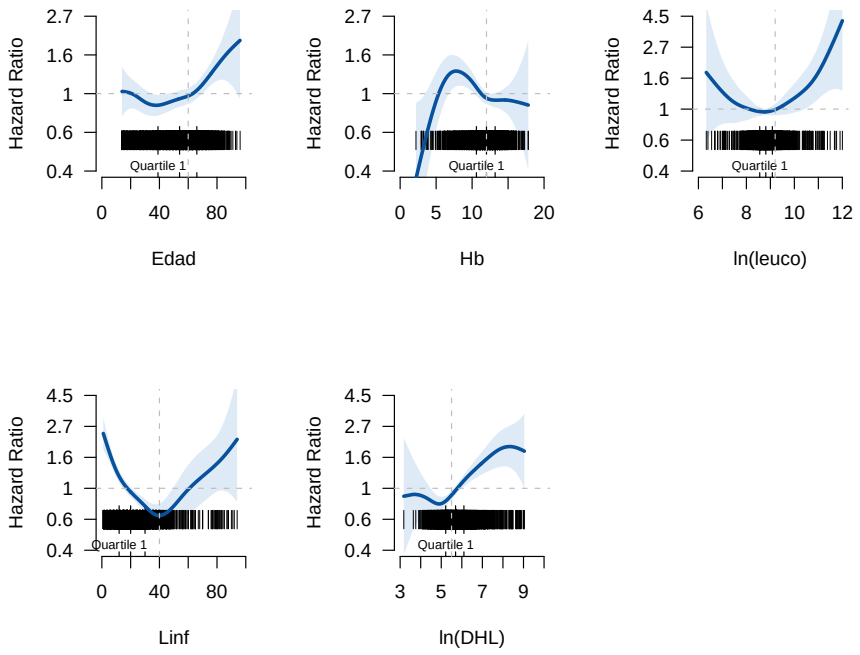
En los resultados (Tabla 4) se observa que los factores pronósticos con efecto significativo ($p < 0.05$) para la supervivencia global de los pacientes con LNH fueron casi todas las variables, a excepción del primario ($p = 0.190$) y los síntomas ($p = 0.140$). Sin bien la edad, el zubrod, el estadio clínico (EC) y la deshidrogenasa láctica (DHL) son factores pronósticos muy conocidos y reportados en esta patología, aquí además se identifican la hemoglobina (Hb), número de leucocitos y el porcentaje de linfocitos con efectos significativos ($p < 0.05$) para la supervivencia global en los LNH.

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 tasa de riesgo, debido a que el efecto de las covariables no presenta una relación lineal con la función de riesgo. En los resultados (Tabla 2) se observa que todas las variables continuas (edad, Hb, leucocitos, linfocitos y DHL) presentan un efecto no lineal significativo en la función de riesgo.

En la Gráfica 11 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.

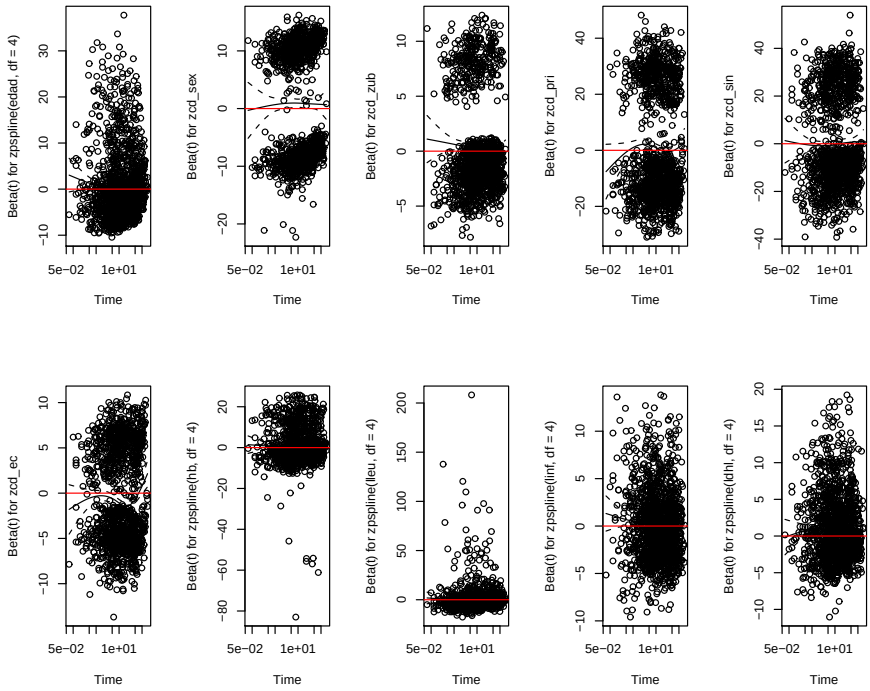


Gráfica 11. Forma funcional de la razón de riesgo mediante el modelo de Cox con P-spline.

Para el número de leucocitos, el riesgo de mortalidad para los leucocitos entre 3-10 mil es menor a HR=1; sin embargo, 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 deshidrogenasa láctica (DHL), el riesgo de mortalidad para un nivel de DHL menor a 240 UI/L es aproximadamente HR=1, sin embargo, cuando el nivel de DHL se incrementa después de 240 UI/L el riesgo se incrementa.

En la Gráfica 12 se muestra los residuos de Schoenfeld escalados vs. tiempo, para las variables categóricas (género, zubrod, primario, estadio clínico y síntomas). En los cuales, se observa que las curva de ajuste se aproximan a una línea recta horizontal, lo que sugiere el cumplimiento del supuesto de riesgo proporcional constante.



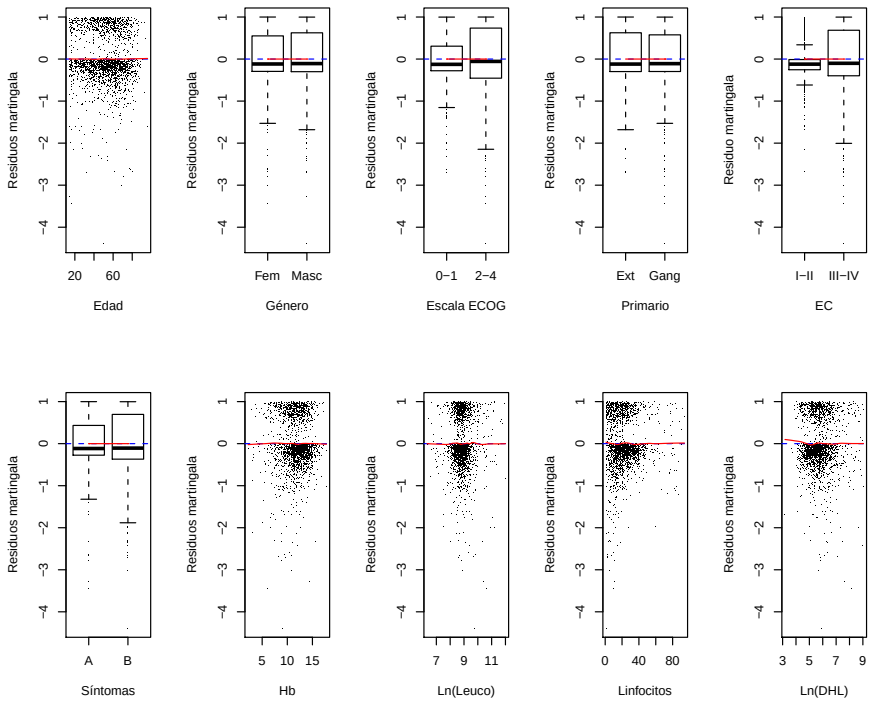
GRÁFICA 12. Residuos de Schoenfeld escalados vs. tiempo ($ln(meses)$) del modelo de Cox con P-spline. En la primera fila de izquierda derecha se muestra para la edad, género, zubrod, primario y estadio clíncio. En la segunda fila de izquierda a derecha se muestra para síntomas, Hb, ln(leucocitos), linfocitos y ln(DHL)

En la Tabla 5 se muestran los resultados del test de no proporcionalidad de Therneau y Grambsch. En los resultados se observan que el efecto de las covariables es constante (Test de no proporcionalidad global: $p > 0.05$). Por lo tanto, los resultados de las estimaciones bajo el modelo de Cox con P-splines son válidos.

En la Gráfica 13 se muestran los residuos martingala para el modelo de Cox con P-splines. Los resultados indican que los residuos para cada covariable son constantes, lo cual verifica que los efectos de las covariables son bien aproximados por los P-splines.

TABLA 5. Prueba individual y global para la proporcionalidad, basado en el residual de Schoenfeld escalado del modelo de Cox con P-splines. La correlación de Pearson es entre los residuos y $g(t)$ para cada covariable.

Variables	correlación de Pearson (rho)	χ^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 13. Residuos martingala del modelo de Cox con P-spline.

5.4. Aplicando el modelo de Cox con PF

En la Tabla 6 se muestra los resultados del modelo de Cox con polinomio fraccional para aproximar el efecto de las covariables continuas (edad, Hb, ln(leuco), linfocitos y ln(DHL) en la función de riesgo, incluyéndose las covariables categóricas como el género, zubrod, primario, estadio clínico y síntomas.

Tabla 6. Resultados del modelo de Cox con polinomio fraccional.

Variables	$\hat{\beta}$	EE($\hat{\beta}$)	Z	p	HR (IC95 %)
Edad: $I((edad/100)^3)$	0.765	0.248	3.086	0.002	2.15 (1.32, 3.49)
Género masculino	0.227	0.078	2.926	0.003	1.26 (1.08, 1.46)
Zubrod 2-4	0.631	0.084	7.474	<0.001	1.88 (1.59, 2.22)
Primario ganglionar	-0.087	0.085	-1.023	0.306	0.92 (0.78, 1.08)
EC III-IV	0.402	0.086	4.683	<0.001	1.50 (1.26, 1.77)
Síntomas B	0.110	0.082	1.340	0.180	1.12 (0.95, 1.31)
Hb1: $I((Hb/10)^{-2})$	0.675	0.172	3.915	<0.001	1.96 (1.40, 2.75)
Hb2: $I((Hb/10)^{-2} \times ln(Hb/10))$	0.649	0.164	3.955	<0.001	1.91 (1.39, 2.64)
Ln(leuco): $I((lleuco/10)^1)$	1.159	0.678	1.709	0.087	3.19 (0.84, 12.04)
Linfocitos1: $I((linf/10)^{0,5})$	-1.229	0.140	-8.758	<0.001	0.29 (0.22, 0.39)
Linfocitos2: $I((linf/10)^2)$	0.039	0.006	6.532	<0.001	1.04 (1.03, 1.05)
Ln(DHL): $I(ldhl/10)^1)$	2.600	0.451	5.762	<0.001	13.46 (5.56, 32.58)
LR test	427.20				
AIC	9525.13				

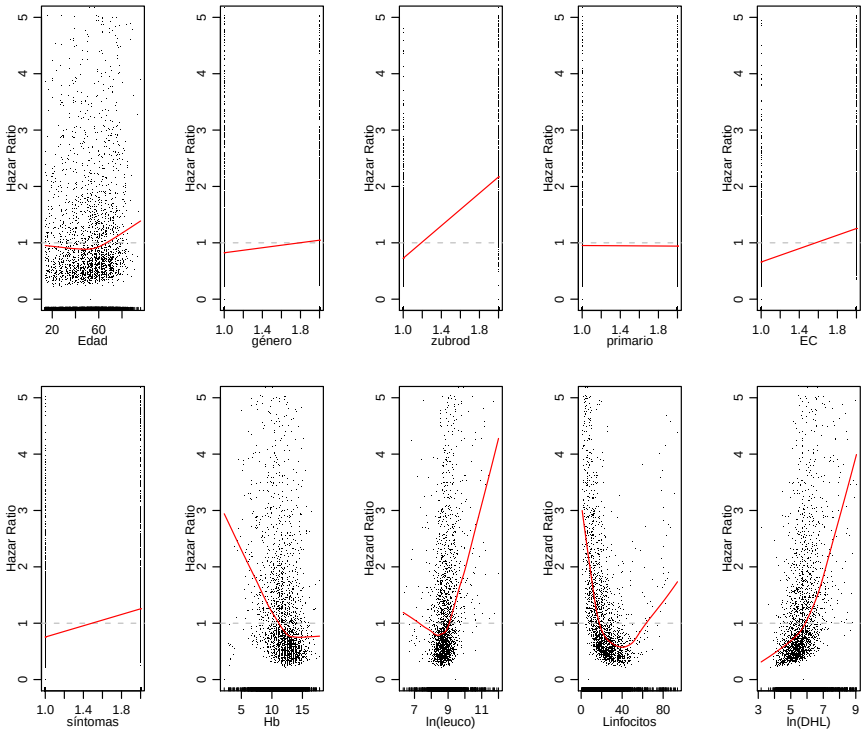
Nota: Las categorías que no aparecen corresponden a las categorías de referencia.
Beta: Coeficiente de regresión. EE: error estándar del coeficiente de regresión.
HR: Hazard ratio.

En los resultados de la Tabla 6 se observa que los factores pronósticos con efecto significativo ($p < 0.05$) en la supervivencia global de los pacientes con LNH son casi todas las variables, a excepción del primario ($p = 0.306$), síntomas ($p = 0.180$) y los leucocito (p=0.087). Los resultados de este modelo son similares al modelo con P-splines a excepción del efecto de los leucocitos.

Para las covariables categóricas la tasa de riesgo de estas variables implican que, los pacientes de sexo masculino presentan un riesgo de mortalidad de HR=1.3 (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).

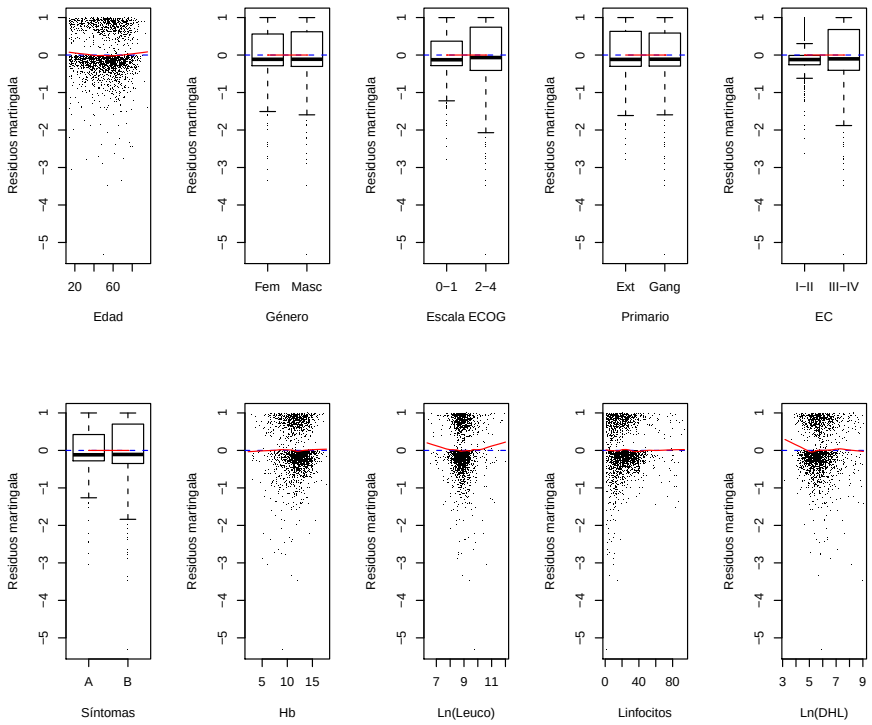
Según el método de polinomio fraccional cada covariable continua requirió un tipo de trasformación. La edad fue dividida por 10 y después requirió una trasformación de potencia cúbica, la Hb fue transformada en $(I(Hb/10))^{-2} + I((Hb/10))^{-2} \times ln(Hb/10)$, el logaritmo del número de leucocitos fue dividida por 10, el porcentaje de linfocitos fue transformada en $I((linfocitos/10))^{0,5} + I((linfocitos/10))^2$ y el logaritmo de DHL fue dividida por 10.

En la Gráfica 14 se muestra la forma funcional de las covariables en la razón de riesgo. Para la edad, el riesgo de mortalidad en menores de 60 años es menor a HR=1, sin embargo, el riesgo se incrementa ligeramente después de los 60 años de edad. Para la Hb $\geq$12g/dl, el riesgo de mortalidad es aproximadamente menor a HR=1, sin embargo, el riesgo se incrementa cuando la Hb disminuye después de 12g/dl. Para el número de leucocitos, el riesgo de mortalidad es aproximadamente menor a HR=1 para menores a 10 mil, y cuando el número de leucocitos se incrementa después de los 10 mil leucocitos el riesgo de mortalidad también se incrementa (efecto de hiperleucocitosis); aunque aquí no se detecta el riesgo de mortalidad por leucopenia como fue identificado por los splines. Para el porcentaje de linfocitos, el riesgo de mortalidad se incrementa cuando el porcentaje de leucocitos disminuyen después 20 %, así mismo, se incrementan después de los 60 %. Para la deshidrogenasa láctica (DHL), el riesgo de mortalidad para un DHL menor a 240UI/L es menor a HR=1, sin embargo, cuando el nivel de DHL se incrementa después de 240UI/L el riesgo también se incrementa.



Gráfica 14. Forma funcional de la razón de riesgo mediante el modelo de Cox con PF.

En el Gráfico 15 se muestra los residuos martingala para el modelo de Cox con polinomio fraccional. Los resultados indican que los residuos para cada covariable son aproximadamente constantes, lo cual verifica que los efectos de las covariables son adecuadamente aproximados utilizando el método de polinomio fraccional.



Gráfica 15. Residuos martingala del modelo de Cox con PF.

5.5. Comparación de los modelos

En la Tabla 7 se muestran los resultados resumidos del ajuste de los tres modelos (Cox clásico, Cox con P-splines y Cox con polinomio fraccional). 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)

Los factores pronósticos identificados mediante el modelo de Cox clásico fueron casi todas las covariables incluidas en el análisis, a excepción del foco primario (p=0.583); sin embargo, en este modelo las variables como zubrod, primario, síntomas, ln(leucocitos) y ln(DHL) no cumplían los supuestos de riesgo proporcional. Así mismo, los residuos martingala muestran que el efecto de estas variables en la supervivencia presentan una forma funcional no lineal, lo que sugería el uso de métodos más flexibles para aproximar el efecto de las covariables en la función de riesgo.

En cambio con el modelo de Cox con P-splines y modelo de Cox con polinomio fraccional las covariables con efecto significativo fueron casi todas a excepción del primario y síntomas ($p > 0.05$). Si bien ambos modelos describen la forma funcional

no lineal de los efectos de las covariables, el ajuste con P-splines describe mejor la forma funcional en algunas covariables que no es identificado por el polinomio fraccional. Un ejemplo, de esto es para logaritmo de leucocitos donde P-splines describe dos grupos de riesgo (para <3mil y >10mil) y el polinomio fraccional solo identifica un grupo de riesgo linealmente creciente para leucocitos mayores de 10mil.

Los residuos martingala para ambos modelos (modelo de Cox con P-splines y modelo de Cox con polinomio fraccional) según las covariables 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 por cada covariable son aproximadamente constantes en consecuencia los ajustes de ambos modelos son adecuados.

Sin embargo, de acuerdo al criterio de información de Akaike (AIC), el modelo de Cox con splines presenta ligeramente una menor AIC (9523.97) en comparación al modelo de Cox con polinomio fraccional (AIC: 9525.13).

Finalmente en la Tabla (8) se muestra los factores pronósticos para la supervivencia de los pacientes con Linfoma no Hadgkin, bajo el modelo de Cox con P-splines. Las variables con efecto significativo para la supervivencia a un nivel de significación de significación de 5 % ($\alpha = 0{,}05$) fueron: la edad ($p = 0.028$), género ($p = 0.009$), zubrod ($p < 0.001$), EC ($p < 0.001$), Hb ($p = 0.006$), leucocitos ($p = 0.038$) y linfocitos ($p < 0.001$), así mismo, la DHL ($p = 0.089$) por ser clínicamente relevante por su significado pronóstico en los LNH.

TABLA 7. Resultados del modelo de Cox con polinomio fraccional.

	Modelo de Cox			Modelo de Cox con P-spline			Modelo de Cox con PF		
Variables	sin transformación	p	HR	P-spline	p	HR	PF	p	HR
Edad:	$\surd$	0.027	1.01	lineal	0.007		$I((edad/100)^3)$	0.002	2.15
	-	-	-	no-lineal	0.028	pspline(, df=4)	-	-	-
Género masculino	$\surd$	0.009	1.22	$\surd$	0.009	1.23	$\surd$	0.003	1.26
Zubrod 2-4	$\surd$	<0.001	1.99	$\surd$	<0.001	1.86	$\surd$	<0.001	1.88
Primario ganglionar	$\surd$	0.583	0.95	$\surd$	0.109	0.89	$\surd$	0.306	0.92
EC III-IV	$\surd$	<0.001	1.55	$\surd$	<0.001	1.49	$\surd$	<0.001	1.50
Síntomas B	$\surd$	<0.001	1.18	$\surd$	0.140	1.13	$\surd$	0.180	1.12
Hb:	$\surd$	0.002	0.95	lineal	0.082		$I((Hb/10)^2)$	<0.001	1.96
	-	-	-	no-lineal	0.006	pspline(, df=4)	$I((Hb/10)^{-2} \times ln(Hb/10))$	<0.001	1.91
ln(Leucocitos):	$\surd$	0.002	1.24	lineal	0.082		$I(ln(leuco)/10)$	0.087	3.19
	-	-	-	no-lineal	0.038	pspline(, df=4)	-	-	-
Linfocitos:	$\surd$	<0.001	0.99	lineal	<0.001		$I((linf/10)^{0,5})$	<0.001	0.29
	-	-	-	no-lineal	<0.001	pspline(, df=4)	$I((linf/10)^2)$	<0.001	1.04
ln(DHL):	$\surd$	<0.001	1.29	lineal	<0.001		$I(ln(DHL/10))$	<0.001	13.46
	-	-	-	no-lineal	0.089	pspline(, df=4)	-	-	-
LR test	351.00			455.00			427.2		
AIC	9597.33			9523.97			9525.13		

Nota: Las categorías: *ln*: *logaritmo natural*, $\surd$: sin trasformación.

TABLA 8. Factores pronósticos para 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)
LR test		455.00
AIC		9523.97

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 supervivencia ha contribuido al desarrollo de métodos más sofisticados desde técnicas simples a más complejas, las cuales han crecido rápidamente durante los últimos años para un mejor modelamiento, facilitado por el rápido 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: riesgo proporcional, 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 cumple, 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 odds proporcional y modelo log-logístico o modelo de Cox ponderado podrían ser una alternativa, pero se debe tener en cuenta en todos estos modelos la forma lineal del efecto de las covariables.

En este trabajo, se utilizaron métodos más flexibles como son: método de suavizamiento P-spline y polinomio fraccional, 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 P-splines y el modelo de Cox con polinomio fraccional 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 continua presenta una estructura de relación cuya forma funcional es no lineal.

Los factores pronósticos con efecto significativo para la supervivencia en LNH fueron: la edad, género, zubrod, estadio clínico (EC), nivel de hemoglobina (Hb), leucocitos, linfocitos y la deshidrogenasa láctica (DHL) como en el modelo de Cox con P-splines y el modelo de Cox con polinomio fraccional. Los cuales, concuerdan con

los reportados en la literatura para esta patología (Nicolaides, Dimos y Pavlidis, 1998; Rebasa, 2001)

El modelo de Cox con P-splines y el modelo de Cox con polinomio fraccional aproximan bien el efecto de las covariables, sin embargo, el método basado en los P-splines describe mejor la forma funcional de la razón de riesgo para las covariables continuas que el modelo de Cox con polinomio fraccional, siendo esta una muy buena alternativa en situaciones donde la forma funcional es no lineal.

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. 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 grupos de mayor riesgo. Para la Hb baja ($<12g/dl$) y los valores elevados de la DHL ($>240U/L$) los puntos de corte obtenidos coinciden con los puntos de corte definidos clínicamente para un peor pronóstico.

Sin embargo, para los valores de los leucocitos y los linfocitos existen dos puntos de corte 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 pronostico deberían ser considerados en la práctica clínica al momento de clasificar a los pacientes en grupos de pronóstico.

Finalmente, este trabajo tiene algunas limitaciones en cuanto a la base 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 las variables como tipo histológico, grados de agresividad e inmunofenotivo (tipo celular). Así mismo, datos como las β-2 microglobulinas que junto con tipo celular son factores pronósticos en este grupo de pacientes.

Como trabajo posterior desde el aspecto clínico, se podría plantear realizar el análisis de los factores pronósticos en los linfomas agresivos, principalmente linfomas de células grandes B difuso, que según la clasifcación de la Organización Mundial de la Salud (clasificación actual de los LNH) representa el 80 % de los LNH. Desde el aspecto metodológico se podría plantear realizar un análisis de los factores pronósticos utilizando el modelo de Cox con splines penalizados para aproximar la forma funcional del efecto de las covariables, así como aproximar los coeficientes tiempo-dependiente para las variables que no son constantes en el tiempo.

Bibliografía

[1] 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.

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

[3] 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.

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

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

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

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

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

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

[10] 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.

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

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

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

[14] 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.

[15] 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.

[16] Keele, L. (2010) Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models. Political Analysis, Vol. 18, pp. 189-205.
[17] Kyle, F. and Hill, M. (2010) NHL (diffuse large B-cell lymphoma). Clinical Evidence, 2010;11:2401
[18] Klein, J.P. and Moeschberger, M.L. (1997) Survival analysis: Technique for censored and truncated data. Springer-Verlag, New York.
[19] 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.
[20] Martinussen, T. and Scheike, T. (2006) Dynamic regression models for survival data. Springer, New York.
[21] 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.
[22] 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.
[23] Nicolaides, C., Dimous, S. and Pavlidis, N. (1997) Prognostic factor in aggressive non-Hodgkin lymphomas. The Oncologist, Vol.3, pp.189-197.
[24] 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.
[25] 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.
[26] Programas Nacionales de Control de Cáncer: Políticas y Pautas para la Gestión. OPS, 2004.
[27] 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.
[28] Royston, P. and Sauerbrei, W. (2008) Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley.
[29] 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.
[30] 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.
[31] Sauerbrei, W., Meier-Hirmer, C., Benner, A. and Royston, P. (2006) Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. Computational Statistical & Data Analysis, Vol. 50, pp. 3464-3485.
[32] 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.
[33] Stone, C.J. (1985) Additive regression and other nonparametric models. The Annals of Statistics, Vol.13, pp. 689-705.
[34] Therneau, T.M. and Grambsch, P.M. (2000) Modeling survival data: extending the Cox model. Springer-Verlag, New York.

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