UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Una aplicación de análisis de supervivencia con variables dependientes del tiempo a un estudio de obesidad.

T E S I S

Que para obtener el título de:

Matemático

P R E S E N T A :

Álvarez Díaz Angel Geovanni

DIRECTOR DE TESIS:
M. en C. José Salvador Zamora Muñoz

Ciudad Universitaria, CD. MX., 2018

Agradecimientos

A mis padres, Leticia y Angel, por ser el pilar fundamental en todo lo que soy, por su incondicional apoyo perfectamente mantenido a través del tiempo. Todo este trabajo ha sido posible gracias a ustedes.

A mi director de tesis M. en C. José Salvador Zamora Muñoz por su gran apoyo y dedicación para la culminación de este trabajo; a la Dra. Leticia Hernández por su tiempo compartido y apoyo ofrecido a este proyecto.

A mis hermanos, Naomi y Brian, por estar conmigo y apoyarme siempre. Espero ser un buen ejemplo para ustedes.

A mis abuelos, por los consejos y apoyo incondicional, son parte importante en mi vida.

A mis tíos, Mary, Carlos y Liz, por acompañarme durante este arduo camino y compartir conmigo alegrías y fracasos.

A mis primas, Jaque y Sandy, por sus palabras de aliento y por siempre creer en mi.

Gracias a todas las personas que ayudaron directa e indirectamente en la realización de este proyecto.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Introducción

La obesidad y sobrepeso infantil es un problema de suma complejidad e importancia debido al gran número de individuos que la padecen, aunado a la falta de interés de la población en el tema. Dicho problema afecta en diversos ámbitos, se asocia a una mayor probabilidad de muerte y discapacidad en una edad adulta, puesto que es un factor importante en el incremento de la probabilidad de padecer una enfermedad cardiovascular. De igual manera interfiere en demasía en la aceptación social, ya que las personas con estas características tienden a ser blanco de burla e insultos por parte del resto de la población y esto puede afecta su autoestima.

Actualmente la obesidad en los adolescentes mexicanos pasó de ser una condición a un problema de primera importancia, ya que México ocupa el sexto lugar en obesidad infantil según la Organización para la Cooperación y Desarrollo Económico (OCDE). La obesidad afecta a todos los grupos de edad y si se pretende controlarla de manera efectiva en los adultos, se requiere prevenirla en la niñez. Las dietas poco equilibradas y el sedentarismo son factores determinantes para su desarrollo y al mismo tiempo son los más difíciles de controlar por el estilo de vida actual.

Si bien no aparece como causa de muerte en los certificados de defunción, la obesidad es un factor importante de riesgo para desarrollar distintas enfermedades de alta mortalidad, tales como: enfermedades coronarias, enfermedades cerebrovasculares y diabetes mellitus.

Se propone realizar un estudio para analizar el tiempo que requiere un joven del Hos-

pital Infantil de México para reducir el 5 % de su índice de masa corporal estandarizado, además de observar qué factores y hábitos alimenticios influyen en el tiempo requerido. Biológicamente la disminución del 5 % es significativa, ya que para una persona con un índice de masa corporal estandarizado muy por encima del recomendado le tomará demasiado tiempo regresar al ideal, además que al disminuir el 5 % el individuo empieza a notar cambios en su físico y salud. Para ello se propone realizar un análisis utilizando el modelo de supervivencia, ya que el objetivo es conocer el tiempo medio requerido para disminuir el 5 % del índice de masa corporal.

La información a utilizar fue recabada por los doctores del hospital infantil de México, "Federico Gómez". El estudio constó de 269 jóvenes que rondan entre los nueve y veintitrés años, de los cuales 119 son mujeres y 150 hombres, además de pertenecer en su mayoría al área metropolitana.

La base de datos que se analizará consta de 1145 registros, ya que cada uno de ellos corresponde a las diversas visitas que hicieron los 269 pacientes al hospital para registrar los cambios en los hábitos alimenticios, edad, talla y peso, además incluye algunas características del sujeto tales como: género, peso al nacer y actividad física.

Para el análisis se contabilizó el número de ocasiones al día y las veces por semana que ingieren ciertos alimentos, a partir de ello se creó un vector alimenticio con los datos del cuestionario aplicado en cada visita de los jóvenes, dicho vector contiene la ingesta total promedio por día de agua, carbohidratos, lípidos, calorías, azúcares, hierro, entro otros nutrientes. Además de utilizar el vector de nutrición se observaron algunas características del individuo como el género, edad y peso al nacer el cual es una característica importante puesto que es uno de los factores que influyen en el riesgo que un adolescente padezca sobrepeso u obesidad.

1.2. Planteamiento del problema

¿Qué factores influyen en el tiempo medio requerido para que un adolescente del Hospital Infantil disminuya el cinco por ciento de su índice de masa corporal estandarizado?

1.2.1. Importancia de la investigación

La obesidad en la adolescencia ha pasado de ser una condición ha ser un problema, puesto que el aumento de adolescentes obesos a incrementado de manera importante, lo cual afectará en años siguientes de manera significativa ya que habrá una población adulta con distintas enfermedades crónicas, cuya causa principal o factor de riesgo principal sea la obesidad. Por lo anterior es importante identificar qué factores influyen en el tiempo que una persona necesita para disminuir su peso corporal, ya que esto permitirá controlar o dar solución al problema de obesidad.

Los factores principales a analizar son los nutrientes que se ingieren con mayor frecuencia, ya que esto permitirá elaborar dietas ricas en nutrientes que favorecen la disminución del peso corporal, además de identificar qué nutrimentos aumentan el riesgo de padecer obesidad.

1.3. Objetivos

General

- Analizar el tiempo medio que requiere para disminuir cinco por ciento de su índice de masa corporal estandarizado, un adolescente del Hospital Infantil de México, "Federico Gómez".

Específicos

- Identificar qué nutrientes y características físicas influyen en que un individuo disminuya su índice de masa corporal.

Capítulo 2

Nutrición

2.1. Macronutrientes

Los macronutrientes son "nutrimentos que cumplen con funciones energéticas y que se encuentran en forma de polímeros y por lo tanto, deben de ser digeridos para que el organismo los pueda utilizar." (Otero Lamas, 2012, p.13)

Los macronutrientes se dividen en tres grupos: carbohidratos, proteínas y lípidos, estos nutrientes componen aproximadamente el 99 por ciento de la dieta de una persona.

Carbohidratos

Los carbohidratos son la principal fuente de energía en la dieta de una persona puesto que se caracterizan por ser unidades estructurales de azúcar. Podemos dividir los hidratos de carbono en tres grupos: monosacáridos, disacáridos y polisacáridos.

Los monosacáridos son los componentes básicos de los azúcares, los más importantes en la dieta humana son la glucosa, la fructosa y la galactosa. La glucosa es la molécula de azúcar más importante puesto que es necesaria para el óptimo funcionamiento del cerebro, se obtiene de la mayoría de alimentos esencialmente de los que tiene un sabor dulce. Por otro lado la fructosa, tal como su nombre indica, es el azúcar natural de las frutas. Los disacáridos son la combinación de dos monosacáridos, existen una gran variedad de disacáridos por lo cual nos basaremos en los tres más importantes: sacarosa, lactosa y maltosa.

"La sacarosa (azúcar de mesa, azúcar de caña) se forma cuando se unen entre sí la glucosa y la fructosa." (Otero Lamas, 2012, p.14)

La lactosa o azúcar de la leche está constituida por glucosa y galactosa, la cual es sintetizada principalmente en las glándulas mamarias de los animales. Durante el proceso digestivo se produce la galactosa a partir de la lactosa.

La maltosa o azúcar de malta está formada por dos moléculas de glucosa, es difícil encontrar de modo natural en los alimentos, aunque se forma por la hidrólisis de polímeros de almidón y es parte de gran variedad de alimentos (pan, cerveza y cereales) como un aditivo.

Por último hablemos de los polisacáridos, los cuales son carbohidratos con más de diez unidades de monosacáridos; asimismo nos enfocaremos en la fibra ya que es el polisacárido de mayor ingesta.

La fibra alimentaria se puede definir como la parte comestible de las plantas que resiste la digestión y absorción en el intestino delgado y que experimenta una fermentación parcial o total en el intestino grueso, se puede encontrar principalmente en los vegetales y cáscaras de algunas frutas. Existen dos tipos de fibra: la dietética o insoluble está en granos integrales, verduras y salvado de trigo. Este tipo de fibra ayuda al paso de los alimentos por los intestinos. El otro tipo es la fibra funcional o soluble la cual atrae el agua y hace que el proceso digestivo sea lento, además de reducir el colesterol. Este tipo de fibra se encuentra en alimento como nueces, semillas, lentejas y algunas frutas y verduras.

Podemos concluir que los hidratos de carbono son la mayor parte de nutrientes que aportan la energía que se necesita, así como la ingesta de dichos nutrientes depende de la actividad física de cada persona. Esto quiere decir que si la persona mantiene una vida sedentaria la ingesta de carbohidratos debe disminuir. Además entre el 60 % y 65 % del total de calorías deben ser aportadas por la ingesta de carbohidratos.

Lípidos

Los lípidos son unas moléculas compuestas de ácidos grasos, aproximadamente entre el 20 a 25 % de la energía de la dieta humana es aportada por ellos. Las grasas son parte esencial para la digestión, absorción y transporte de las vitaminas liposolubles además rodean los órganos protegiéndolos de golpes y traumas.

Los lípidos se clasifican por el número de enlaces y posición de éstos en:

a) Lípidos simples: Son ácidos grasos que pueden ser saturados, como los que contienen la manteca, la mantequilla, la crema; monoinsaturados, como los del aceite de maíz, girasol o de la nuez de macadamia; y poliinsaturados, como los que contienen los cacahuates, las nueces o el aceite de oliva.

b) Lípidos compuestos: Son los fosfolípidos que se encuentra en el hígado y el huevo; los glucolípidos que existen en alimentos de origen animal tales como la leche y el atún; y las lipoproteínas que contiene la piel del pollo y la margarina.

c) Lípidos misceláneos: Son las vitaminas A,K y E y los esteroles. Principalmente se encuentran en alimentos de origen animal.

Un lípido de vital importancia es el colesterol. El colesterol lo fabrica el cuerpo humano, además se puede encontrar en los alimentos de origen animal. Este tipo de grasa es necesario para fabricar las membranas celulares y algunas hormonas como el estrógeno. El consumo ideal de colesterol se debe limitar a menos de 300 mg por día ya que su consumo en exceso aumenta el riesgo de padecer una enfermedad del corazón.

Proteínas

En comparación con los lípidos e hidratos de carbono las proteínas son el macronutriente que aporta menor energía al organismo, pero son necesarias para otras funciones importantes. Las proteínas constituyen la base para construir los tejidos del cuerpo (piel, músculos, sangre, huesos), especialmente en los periodos de crecimiento, además de ayudar a repararlo. Como se mencionó anteriormente las proteínas aportan energía, un gramo equivale a 4 kilocalorías.

Podemos encontrar las proteínas principalmente en lo alimentos de origen animal, especialmente en la carne de los animales. Aunque los alimentos de origen animal cubren más fácilmente las necesidades de proteínas, existen alimentos de origen vegetal que de igual forma contribuyen a la ingesta de proteínas, tales como legumbres, semillas (nueces, almendras, cacahuates), pan y pastas.

Las cantidades necesarias de proteínas para el cuerpo humano dependen del peso, género y edad del individuo. En la tabla 1.1 se muestra la cantidad recomendada que cada individuo debe consumir de proteínas al día. Las cantidades están en gramos de proteína a ingerir por kilogramos que pese el individuo.

Tabla 2.1: Cantidad diaria recomendada de proteínas

Género	Edad	Ingesta Recomendada, g/kg/día
Niños	5 - 12 años	1.35
Hombres	12 - 14 años	1.35
	14 - 16 años	1.3
	16 - 18 años	1.2
	18 y más años	1.0
Mujeres	12 - 14 años	1.3
	14 - 16 años	1.2
	16 - 18 años	1.1
	18 y más años	1.0

Fuente: Guías de alimentación. Bases para su desarrollo en América Latina. Reunión UNU/Fundación CAVENDES.Caracas 1988.

2.2. Micronutrientes

En la sección anterior hablamos de los nutrientes que aportan en mayor medida energía y otras características dentro de una dieta, ahora hablemos sobre los nutrientes que se encuentran en menor porción dentro de los alimentos, pero esto no quiere decir que sean menos importantes. Los micronutrientes son las vitaminas, minerales o nutrimentos inorgánicos.

Vitaminas

Las vitaminas son un grupo de micronutrientes esenciales que cumplen los siguientes criterios:

1. Componentes que se encuentran en los alimentos habitualmente en cantidades muy pequeñas.
2. No producidos o sintetizadas por el cuerpo en cantidades suficientes para satisfacer las necesidades del mismo.
3. Esenciales, también en cantidades pequeñas, para una función fisiológica normal (es decir, crecimientos, desarrollo y reproducción).

Las vitaminas se clasifican de la siguiente manera:

a) Vitaminas liposolubles: Son la A,D,E y K.

b) Vitaminas hidrosolubles: Son el ácido pantoténico, niancina, riboflavina o B2, ácido fólico, B12, B6, B1 y ácido ascórbico o vitamina C.

En la siguientes tablas se da una breve explicación sobre la función y los alimentos que contienen ciertas vitaminas.

Tabla 2.2: Función y fuentes de vitaminas hidrosolubles.

Vitamina	Función	Fuentes
Ácido pantoténico	Transferencia de grupos acilo y acetilo	Todos los alimentos
Niancina	Reacciones de óxido-reducción	Tejidos de animales, tortilla y leche
Vitamina B2	Reacciones de óxido-reducción	Tejidos de animales, huevo y leche
Ácido fólico	Metabolismo en un solo carbón	Hojas verdes y víceras
Vitamina B6	Reacciones de metilación	Flora intestinal, leche y tejidos animales
Vitamina B1	Reacciones de descarboxilación	Semillas maduras de cereales
Vitamina C	Reacciones de carboxilación y transcarboxilación	Tejidos vegetales frescos

Nota. Fuente: Otero Lamas.(2012).*Nutrición* (p.19)

Tabla 2.3: Función y fuentes de vitaminas liposolubles.

Vitamina	Función	Fuentes
Vitamina A	Ciclo visual, diferenciación celular y respuesta inmune	Tejidos animales y leches
Vitamina E	Antioxidante	Aceites vegetales
Vitamina K	Factor de coagulación y la calcificación ósea	Hojas verdes y flora intestinal
Vitamina D	Absorción y metabolismo del calcio, mineralización, contracción muscular y respuesta inmune	Tejidos animales, especialmente hígado. En presencia de luz ultravioleta, síntesis en la piel

Nota. Fuente: Otero Lamas.(2012).*Nutrición* (p.20)

2.3. Obesidad

La Organización Mundial de la Salud (OMS) define a la obesidad y sobrepeso como ”una acumulación anormal o excesiva de grasa que puede ser perjudicial para la salud". El sobrepeso y obesidad son factores de riesgo para numerosas enfermedades crónicas y aunque no aparezca como causa de la muerte de una persona, es un factor importante para ello. En ocasiones se piensa que el riesgo de padecer obesidad es mayor para los adultos, por la vida sedentaria que llevan, pero esto no es así, la obesidad infantil es uno de los principales problemas de salud publica mundial y está afectando no solo a los países con altos ingresos, sino también a los de medianos y bajos, principalmente en las áreas urbanas. Los niños obesos y con sobrepeso tienden a seguir siendo obesos en

una edad adulta, además tienen mayor probabilidad de padecer, a una edad temprana, enfermedades crónicas.

La principal causa de obesidad es el desequilibrio entre la cantidad de calorías ingeridas y el gasto calórico, por lo cual el número de individuos obesos aumenta día con día, ya que actualmente las personas consumen alimentos hipercalóricos excedentes en grasas aunado al estilo de vida sedentario con poca o nula actividad física.

Pero ¿cómo podemos saber si padecemos obesidad o sobrepeso? Esto es simple, existe un indicador que nos permite saber si tenemos sobrepeso o no, este indicador es el índice de masa corporal (IMC), y para ser calculado se necesitan sólo dos medidas, el peso y la estatura de la persona. El IMC se calcula de la manera siguiente

$$IMC = \frac{Peso}{(estatura)^2}$$

donde el peso debe ser en kilogramos y la estatura en metros.

La OMS propone una clasificación según el IMC para el grado de obesidad en los adultos la cual se muestra en la tabla siguiente

Tabla 2.4: Clasificación del sobrepeso y obesidad en adultos de acuerdo con el índice de masa corporal, según los criterios de la OMS (Organización Mundial de la Salud).

Clasificación	IMC	Riesgo de comorbilidades
Bajo peso	<18.5	Bajo (pero con riesgo de otros problemas de salud)
Peso normal	18.5 - 24.9	Promedio
Sobrepeso	25 - 29.9	
Obeso	$\geq$ 30	
Grado I	30 - 34.9	Moderado
Grado II	35 - 39.9	Importante
Grado III	$\geq$ 40	Muy importante

Nota. Fuente: Otero Lamas.(2012).*Nutrición* (p.20)

El índice de masa corporal es un excelente indicador para conocer el grado de obesidad aunque presenta un inconveniente importante el cual es que no puede ser utilizado en personas menores de 19 años, puesto que aún siguen en desarrollo. La OMS propone ciertos criterios sobre cómo saber si una persona menor de 19 años y mayor de cinco padece obesidad y sobrepeso, éstos dependen de la edad y estatura de la persona, además la OMS fija una media del IMC por edad y sexo, por lo cual se dice que una niña o niño tiene sobrepeso si está por encima de la mediana dos desviaciones típicas de la misma, asimismo padece obesidad si está por encima de las tres desviaciones típicas de la mediana establecida en los patrones de crecimientos dados por la OMS.

El tratamiento para un paciente con obesidad tiene como objetivo disminuir su peso

para mejorar o recuperar su salud. Para ello es de suma importancia que disminuya su ingesta caloríca mediante un plan de alimentación adecuado, además de cambiar su estilo de vida sedentario, es decir, que incremente el tiempo que dedica a actividades físicas. Depende cúal sea el grado de obesidad del sujeto para considerar si es necesario suministrar algún fármaco que lo ayude a disminuir peso o si es necesario recurrir a cirugía bariátrica esto en casos donde el índice de masa corporal sea mayor a cuarenta.

2.3.1. IMC estandarizado o z-score

Puesto que los adolescentes están en constante desarrollo físico no podemos utilizar el IMC como indicador de obesidad, por lo cual se recomienda realizar una estandarización del mismo, la cual se calcula de la manera siguiente

$$Z = \frac{IMC - Mediana\ del\ IMC}{Desviación\ estándar}$$

donde la mediana depende del sexo y edad al igual que la desviación estándar, estos valores los ofrece la OMS. Al igual que el IMC existen rangos en los cuales podemos decir que un adolescente tiene sobrepeso u obesidad, tales rangos son:

- Si $Z > 1$ la persona tiene sobrepeso.
- Si $Z > 2$ la persona tiene obesidad.

Capítulo 3

Análisis de supervivencia

3.1. Introducción

El análisis de supervivencia es un conjunto de métodos estadísticos que se utilizan para analizar datos, tales datos se caracterizan por tener un momento de inicio y uno final de observación, al tiempo transcurrido entre estos dos momentos se le denomina tiempo de supervivencia.

Los estudios de supervivencia son muy comunes en el área biomédica, en los cuales el evento de interés es el tiempo para la muerte de un paciente o recuperación de una enfermedad, pero pueden presentarse en diversas áreas del conocimiento. Por ejemplo, en economía podemos dar seguimiento al tiempo de desempleo o de empleo de una persona económicamente activa, en educación se puede seguir el tiempo de deserción de los alumnos en una escuela.

Las principales característica de los datos de supervivencia son las siguientes: tienen un tiempo de supervivencia, el cual se define como el tiempo transcurrido desde el inicio del estudio hasta que el individuo presenta el evento de interés o falla, cabe destacar que el tiempo de falla es una variable aleatoria positiva, generalmente con sesgo positivo, por lo cual se descarta el uso de la distribución normal como un modelo recomendable. Aunque esta característica de positividad del tiempo de supervivencia puede eliminarse si se utiliza una transformación de esta variables, por ejemplo una transformación logarítmica; sin embargo es recomendable trabajar con la variables en su escala natural.

Por otro lado existe la presencia de un tipo especial de datos los cuales caracterizan al análisis de supervivencia, estos son los llamados datos censurados. Puesto que se

trata de estudios longitudinales es común que algunos individuos abandonen el estudio antes de presentar la falla, por lo cual se tendría información parcial de su tiempo de falla a estos individuos se les conoce como observaciones o datos censurados.

Existen diferentes maneras en las cuales se pueden presentar la censura, por lo cual existen distintas clasificaciones.

La censura tipo I se presenta cuando se fija un tiempo máximo para el estudio y algunos individuos no presentan la falla antes del periodo máximo. Estos individuos constituyen las observaciones censuradas.

La censura tipo II se presenta cuando se fija una cantidad finita de fallas, es decir, hasta que se presenten k fallas de n posibles. Los individuos que no experimentaron la falla al completarse las primeras k, representan observaciones censuradas.

La censura aleatoria ocurre sin ningún control de la persona que lleva a cabo el estudio, por ejemplo que el individuo abandone el estudio, o muerte por alguna causa que no está relacionada con el estudio.

Puesto que se trata de modelar una variable aleatoria T no negativa la cual representa el tiempo de falla de los sujetos en el estudio, para ello existen cuatro funciones que nos ofrecen información de la variable aleatoria.

1. Función de supervivencia

   La función de supervivencia representa la probabilidad de que un sujeto no presente la falla (sobreviva) en un tiempo t. Es decir $S(t) = \mathbb{P}(T > t)$, a partir de dicha definición podemos ver la función de supervivencia de otro modo $S(t) = 1 - F(t)$ donde $F(t)$ es la función de distribución de T. Las principales características de la función de supervivencia son:

   a) $S(0) = 1$

   b) $S(t) = 0 \; si \; t \to \infty$

   c) $S(t_1) \geq S(t_2) \; si \; t_1 < t_2$

   Además si T es continua, con función de densidad $f(t)$, entonces

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

Para el caso en el cual T es discreta con función de masa de probabilidad $f(T = t_j),\ j = 1, 2 \ldots$ y $t_1 < t_2 < \ldots$ Entonces, su función de supervivencia es

$$S(t) = \mathbb{P}(T > t) = \sum_{t_j > t} f(t_j)$$

2. Función de riesgo

   Determina la tasa instantánea de falla al tiempo $T = t$, dado que el sujeto ha sobrevivido un instante antes de t, y se define de la siguiente forma

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\left(t < T \leq t + \Delta t | T > t\right)}{\Delta t}$$

   y para el caso discreto

$$h(t_i) = \mathbb{P}\left(T = t_i | T \geq t_i\right), \quad i = 1, 2, \ldots$$

   Esta función puede servir de guía para proponer un modelo paramétrico par el tiempo de supervivencia.

3. Función de riesgo acumulado

   En el caso continuo dicha función se define como

$$H(t) = \int_0^t h(u)\ du$$

   y el discreto

$$H(t) = \sum_{t_i \leq t} h(t_i)$$

   además cumple con las siguientes propiedades.

   a) $H(0) = 0$

   b) $si\ t_1 > t_2\ entonces\ H(t_1) \geq H(t_2)$

4. Función de vida media residual

   Por último hablemos de la función de vida media residual la cual se denota como $mrl(t)$, esta función mide la esperanza de vida restante para sujetos con tiempo t, es

decir, el tiempo esperado de vida después de t, hasta la ocurrencia de la falla. Por lo cual se define de la siguiente manera

$$mrl(t) = \mathbb{E}\ (T - t|T > t)$$

Existen algunas relaciones entre las funciones anteriores, comencemos con la función de supervivencia

$$S(t) = 1 - F(t) \Rightarrow\ f(t) = -\frac{dS(t)}{dt}$$

Ahora sigamos con la definición de la función de riesgo $h(t)$

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}\ (t < T \leq t + \Delta t|T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t\ \mathbb{P}\ (T > t)} \\ &= -\lim_{\Delta t \to 0} \frac{1 - F(t + \Delta t) - (1 - F(t))}{\Delta t}\ \frac{1}{S(t)} \\ &= \lim_{\Delta t \to 0} \frac{S(t + \Delta t) - S(t)}{\Delta t} * \frac{1}{S(t)} \\ &= -\frac{S(t)'}{S(t)} \\ &= -\frac{dlog(S(t))}{dt} \end{aligned}$$

Partiendo de lo anterior es fácil encontrar una relación entre la función de riesgo y la de supervivencia

$$h(t) = -\frac{dlog(S(t))}{dt} \Rightarrow\ S(t) = \exp\left\{-\int_0^t h(u)\ du\right\}$$

De igual manera para la función de densidad $f(t)$ tenemos

$$h(t) = -\frac{S(t)'}{S(t)} \Rightarrow\ h(t) = \frac{f(t)}{S(t)} \Rightarrow\ f(t) = h(t)S(t) \Rightarrow\ f(t) = h(t) \exp\left\{-\int_0^t h(u)\ du\right\}$$

Asimismo para $H(t)$

$$S(t) = \exp\left\{-\int_0^t h(u)\, du\right\} \Rightarrow\ S(t) = \exp\{-H(t)\} \Rightarrow\ H(t) = -\log(S(t))$$

Ahora veamos las relaciones entre funciones para el caso discreto

De $S(t)$ con las demás

$$S(t) = \mathbb{P}(T > t) = \sum_{t_j > t} f(t_j) \Rightarrow\ S(t_j) = \sum_{t_i > t_j} f(t_i)\ y\ S(t_{j-1}) = \sum_{t_i > t_{j-1}} f(t_i)$$
$$\Rightarrow\ f(t_j) = -\{S(t_j) - S(t_{j-1})\} = S(t_{j-1}) - S(t_j)$$

Para la relación con la función de riesgo discreta, tenemos

$$h(t_j) = \mathbb{P}(T = t_j | T \geq t_j) = \frac{\mathbb{P}(T = t_j)}{\mathbb{P}(T \geq t_j)} = \frac{f(t))}{S(t_{j-1})} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}$$

De la igualdad anterior tenemos que $S(t_j) = [1 - h(t_j)]S(t_{j-1})$. Desarrollando esta expresión para $j = 1, 2, \ldots$ tenemos

$$\begin{aligned}
S(t_1) &= [1 - h(t_1)]S(t_0) = 1 - h(t_1) \\
S(t_2) &= [1 - h(t_2)]S(t_1) = [1 - h(t_2)][1 - h(t_1)] \\
S(t_3) &= [1 - h(t_3)]S(t_2) = [1 - h(t_3)][1 - h(t_2)][1 - h(t_1)] \\
&\vdots
\end{aligned}$$

Finalmente obtenemos

$$S(t) = \prod_{t_j \leq t} [1 - h(t_j)]$$

Asimismo para $f(t)$ tenemos

$$h(t_j) = \frac{f(t_j)}{S(t_{j-1})} \Rightarrow \ f(t_j) = h(t_j)S(t_{j-1})$$
$$\Rightarrow \ f(t_j) = h(t_j)\prod_{k=1}^{j-1}[1-h(t_k)] = \frac{h(t_j)}{1-h(t_j)}\prod_{k=1}^{j}[1-h(t_k)]$$

3.2. Modelos paramétricos

Existen diferente modelos paramétricos que pueden ser útiles para modelar datos de supervivencia, ya que cualquier distribución con dominio en los reales no negativos puede ser un modelo para el tiempo de supervivencia, pero son pocas de las de uso común. En el siguiente cuadro se presenta las distribuciones más habituales

Tabla 3.1: Modelos paramétricos comunes para el tiempo de supervivencia.

Distribución	Función de distribución $f(x)$	Función de riesgo $h(x)$	Función de supervivencia $S(x)$	Media $\mathbb{E}(x)$
Exponencial $\lambda > 0\ x \geq 0$	$\lambda e^{-\lambda x}$	λ	$e^{-\lambda x}$	$\frac{1}{\lambda}$
Weibull $\alpha, \lambda > 0\ x \geq 0$	$\alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}}$	$\alpha\lambda x^{\alpha-1}$	$e^{-\lambda x^{\alpha}}$	$\frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$
Log logística $\alpha, \lambda > 0\ x \geq 0$	$\frac{\alpha x^{\alpha-1}\lambda}{(1+\lambda x^{\alpha})^2}$	$\frac{\alpha x^{\alpha-1}\lambda}{1+\lambda x^{\alpha}}$	$\frac{1}{1+\lambda x^{\alpha}}$	$\frac{\pi Csc(\pi/\alpha)}{\alpha\lambda^{1/\alpha}}\ si\ \alpha > 0$
Gamma $\beta, \lambda > 0\ x \geq 0$	$\frac{\lambda^{\beta}x^{\beta-1}e^{-\lambda x}}{\Gamma(\beta)}$	$\frac{\frac{\lambda^{\beta}x^{\beta-1}e^{-\lambda x}}{\Gamma(\beta)}}{1-GI(\lambda x,\beta)}$	$1 - GI(\lambda x, \beta)$[1]	$\frac{\beta}{\lambda}$
Log normal $\sigma > 0\ x \geq 0$	$\frac{e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$	$\frac{f(x)}{S(x)}$	$1 - \Phi(\frac{log(x)-\mu}{\sigma})$	$e^{(\mu+0.5\sigma^2)}$
Gompertz $\lambda, \alpha > 0\ x \geq 0$	$\lambda\alpha^x e^{-\frac{\lambda}{log(\alpha)}(\alpha^x-1)}$	$\lambda\alpha^x$	$e^{-\frac{\lambda}{log(\alpha)}(\alpha^x-1)}$	
Pareto $\theta, \lambda > 0\ x \geq 0$	$\frac{\theta\lambda^{\theta}}{x^{\theta+1}}$	$\frac{\theta}{x}$	$\frac{\lambda^{\theta}}{x^{\theta}}$	$\frac{\theta\lambda}{\theta-1}\ si\ \theta > 1$

3.3. Modelos no paramétricos

La función de supervivencia desempeña un papel fundamental para el análisis, pero en ocasiones es complicado saber qué función representa los datos a estudiar, es más, es rara la ocasión en que podemos decir que el tiempo de supervivencia se ajusta a un modelo paramétrico, por lo cual se han propuesto diversos estimadores no paramétricos para la función de supervivencia.

[1] $GI(t,\beta) = \frac{\int_0^t u^{\beta-1}e^{-u}\,du}{\Gamma(\beta)}$

3.3.1. Función de supervivencia empírica

En primera instancia recordemos el teorema fundamental de la estadística

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow{c.s} 0 \;\; cuando \;\; n \to \infty$$

con

$$F_n(t) = \frac{\#T_s \leq t}{n}$$

la función de distribución empírica, así que podemos pensar que la función de supervivencia empírica definida como

$$S_n(t) = \frac{\#T_s > t}{n}$$

puede ser un buen estimador de la función de supervivencia, ya que está definida como la probabilidad de seleccionar una persona que aún no haya presentado la falla al tiempo t de las n posibles. La función de supervivencia empírica es escalonada con saltos de tamaño 1/n en caso de que no ocurran empates, Si hay empates (d) que se definen como el número de sujetos que fallan al mismo tiempo, los saltos son de tamaño d/n. Este estimador presenta una desventaja importante, no toma en cuenta si este tiempo fue falla o censura.

3.3.2. Tabla de vida

Otro estimador para la función de supervivencia es el que se obtiene mediante la tabla actuarial de vida, la cual tiene como principal característica que el tiempo de falla se divide en n intervalos.

$$0 = t_0 < t_1 < t_2 < t_3 < \cdots < t_n = \infty$$

En primer lugar definamos las variables que utilizaremos para construir la tabla de vida

- $I_k = [t_{k-1}, t_k) \;\; k = 1, 2, \ldots, n$

- d_k: número de sujetos que presentan la falla en I_k
- c_k: número de sujetos censurados en I_k
- n_k: número de sujetos en riesgo al inicio de I_k

Ahora sí podemos iniciar la construcción del estimador a partir de la tabla de vida, veamos que

$$\begin{aligned} S(t_{k-1}) &= \mathbb{P}(T \geq t_{k-1}) \\ &= \mathbb{P}(T \geq t_0) \times \mathbb{P}(T \geq t_1 | T \geq t_0) \times \mathbb{P}(T \geq t_2 | T \geq t_1) \times \ldots \times \mathbb{P}(T \geq t_{k-1} | T \geq t_{k-2}) \end{aligned}$$

por otro lado tenemos que

$$\mathbb{P}(T \geq t_{k-1} | T \geq t_{k-2}) = 1 - \mathbb{P}(T < t_{k-1} | T \leq t_{k-2})$$

Para calcular las probabilidades anteriores supongamos que las censuras se distribuyen uniformes en el intervalo.

En el caso que no hubiera censura tendríamos

$$\mathbb{P}(T < t_{k-1} | T \geq t_{k-2}) = \frac{d_k}{n_k}$$

Por el supuesto de uniformidad sobre la censura, podemos considerar que los individuos censurados permanecieron en riesgo a la mitad del intervalo. Por lo cual la probabilidad estimada de sufrir la falla en el intervalo $[t_{k-1}, t_k)$ es

$$\mathbb{P}(T < t_{k-1} | T \geq t_{k-2}) = \frac{d_k}{n_k - \frac{c_k}{2}}$$

Con base en el resultado anterior podemos decir que la probabilidad estimada de sobrevivir al intervalo $[t_{k-1}, t_k)$ es

$$\mathbb{P}(T \geq t_{k-1} | T \geq t_{k-2}) = 1 - \frac{d_k}{n_k - \frac{c_k}{2}}$$

Por lo tanto podemos definir la función de supervivencia estimada de la siguiente

manera

$$\widehat{S}(t_{k-1}) = \prod_{i=1}^{k-1} \left(1 - \frac{d_k}{n_k - \frac{c_k}{2}}\right)$$

la cual representa la probabilidad de que un sujeto sobreviva al inicio del intervalo I_k. En el caso general se tendría

$$\widehat{S}(t) = \widehat{S}(t_{k-1}) \;\; si\; t \in I_k, k = 1, 2, 3, \ldots, n$$

3.3.3. Estimador Kaplan-Meier (K-M)

Por último hablaremos del estimador Kaplan-Meier, el cual es el estimador de máxima verosimilitud de la función de supervivencia. Recordemos las características principales del método de máxima verosimilitud para obtener estimadores.

1. Se tiene una muestra aleatoria $t_1, t_2, \ldots, t_n$ con $t_i \backsim f(\cdot, \theta)$.
2. $L(\theta, t)$ función de verosimilitud, la cual engloba la información que aportan sobre θ, las n observaciones de la muestra.
3. $L(\theta, t)$ para muestra fija de t, es función únicamente de θ.

La función de verosimilitud depende de un parámetro (θ) desconocido, por lo cual la idea es encontrar el valor de θ que maximice la probabilidad de haber observado la muestra, t. Para el estimador Kaplan-Meier queremos estimar una función no un parámetro, por lo cual para aplicar el método de máxima verosimilitud nuestro espacio paramétrico es el espacio de funciones y la función de verosimilitud L es un funcional.

Para su construcción supondremos lo siguiente

- k tiempos de falla $t_1, t_2, \ldots, t_k$
- c_j número de censuras entre dos tiempos consecutivos $j = 0, 1, 2, 3, \ldots, k$
- d_k empates en t_k, es decir, tiempos de falla en t_k

El estimador Kaplan-Meier es el caso límite de la tabla de vida, puesto que el número de intervalos crece, en otras palabras, tiende a infinito, lo cual provoca que su tamaño

disminuya. Entonces estos intervalos deberían contener únicamente una falla o todas las fallas empatadas ahí.

$$I_j = [t_j, t_{j+1})$$

En I_j hay c_j censuras en los tiempos $t_{j1}, t_{j2}, \ldots, t_{jc_j}$

Para construir la verosimilitud $L(S)$ debemos conocer cuál es el aporte de los datos, es decir, si es una falla o censura de qué manera influye en $L(S)$. Los individuos aportan a la verosimilitud de dos forma: Si el individuo presentó la falla contribuye con $f(\cdot, \theta)$, en cambio si se censura su contribución sería $S(\cdot, \theta)$, porque se supone que vivió al menos un instante después del valor de censura, entonces contribuye con $S(t_{ij})$ lo cual nos indica que el individuo se censuró en el i-ésimo intervalo y es la j-ésima censura de las c_j totales dentro de ese intervalo.

Tomemos por ejemplo, si el individuo sufrió la falla en t_n, su contribución es $S(t_n^-) - S(t_n)$=probabilidad de presentar la falla en $t_n = f(t_n)$, donde $S(t_n^-) = \lim_{x \to 0^+} S(t_n - x)$.

Por lo tanto la verosimilitud queda de la siguiente manera. La contribución de las fallas sería

$$\prod_{i=0}^{k} [S(t_n^-) - S(t_n)]^{d_i} \quad (independenciaentreintervalosdetiempo)$$

y la de las censuras

$$\prod_{i=0}^{k} \prod_{j=0}^{c_j} S(t_{ij}) \quad (independencia)$$

Por lo cual la verosimilitud sería

$$L(S) = \prod_{i=0}^{k} \left([S(t_n^-) - S(t_n)]^{d_i} \prod_{j=0}^{c_j} S(t_{ij}) \right)$$

Para que maximice la verosimilitud la función $\widehat{S}(t)$ debe ser discontinua en las fallas, continua en los tiempos de censura y constante entre dos tiempos de falla.

Se puede observar que la verosimilitud es parecida a la función de supervivencia

para el caso discreto sólo hay que realizar algunas asignaciones como

$$S(t_j^-) = \prod_{i=0}^{j-1}(1-\alpha_i),\ j = 1, 2, \ldots, k\ ,\ donde\ \alpha_i\ es\ la\ probabilidad\ fallar\ en\ t_i$$

$$S(t_j) = \prod_{i=0}^{j}(1-\alpha_i),\ j = 1, 2, \ldots, k$$

$$entonces\ \ S(t_j^-) - S(t_j) = \left(\prod_{i=0}^{j-1}(1-\alpha_i)\right)(1-(1-\alpha_j)) = \alpha_j \prod_{i=0}^{j-1}(1-\alpha_i)$$

Ahora consideremos la posibilidad de tener d_j empates en cada tiempo t_j. Entonces

$$[S(t_j^-) - S(t_j)]^{d_j} = \alpha_j^{d_j} \prod_{i=0}^{j-1}(1-\alpha_i)^{d_i}$$

por otro lado tenemos

$$S(t_{ji}) = \prod_{m=0}^{j}(1-\alpha_m) = (1-\alpha_j)\prod_{m=0}^{j-1}(1-\alpha_m)$$

Consideremos todas las censuras en el intervalo $[t_j, t_{j+1})$ se tendría que

$$S(t_{ji}) = (1-\alpha_j)^{c_j} \prod_{m=0}^{j-1}(1-\alpha_m)^{c_j}$$

Por lo que podemos concluir que la verosimilitud es

$$L(S) = \prod_{j=0}^{k}\left(\alpha_j^{d_j}(1-\alpha_j)^{c_j}\prod_{m=0}^{j-1}(1-\alpha_m)^{d_j+c_j}\right)$$

Desarrollando la expresión anterior y agrupando

$$L(S) = \alpha_0^{d_0}(1-\alpha_0)^{c_0} \times \alpha_1^{d_1}(1-\alpha_1)^{c_1}(1-\alpha_0)^{c_1+d_1} \times \alpha_2^{d_2}(1-\alpha_2)^{c_2}(1-\alpha_0)^{c_2+d_2}(1-\alpha_1)^{c_2+d_2} \times$$
$$\cdots \times \alpha_k^{d_k}(1-\alpha_k)^{c_k}(1-\alpha_0)^{c_k+d_k}(1-\alpha_1)^{c_k+d_k} \times \cdots \times (1-\alpha_{k-1})^{c_k+d_k}$$

Antes de seguir con el desarrollo recordemos que definimos a n_j como el número de personas en riesgo de sufrir la falla en el intervalo I_j, los cuales serían al tiempo t_j los vivos y no censurados, entonces

$$n_j = \sum_{i=j}^{k}(d_i + c_j)$$

Por lo cual

$$\begin{aligned} L(S) &= \alpha_0^{d_0}\alpha_1^{d_1}\alpha_2^{d_2}\alpha_3^{d_3}\cdots\alpha_k^{d_k}(1-\alpha_0)^{c_0+d_1+c_1+d_2+c_2+\cdots+d_k+c_k} \times (1-\alpha_1)^{c_1+d_2+c_2+d_3+c_3+\cdots+d_k+c_k} \times \\ &\quad \cdots \times (1-\alpha_k)^{c_{k-1}+d_k+c_k}(1-\alpha_k)^{c_k} \\ &= \alpha_0^{d_0}\alpha_1^{d_1}\alpha_2^{d_2}\cdots\alpha_k^{d_k}(1-\alpha_0)^{n_0-c_0}(1-\alpha_1)^{n_1-c_1} \times \cdots \times (1-\alpha_k)^{n_k-c_k} \end{aligned}$$

Por lo tanto la verosimilitud se puede reescribir como

$$L(S) = \prod_{i=0}^{k}\alpha_i^{d_i}(1-\alpha_i)^{n_i-d_i}$$

De la expresión anterior podemos observar que esta verosimilitud se parece a una $Bernoulli(\alpha_i)$. Por lo cual podemos realizar lo siguiente

$$\begin{aligned} \log(L(S)) &= \log\left(\prod_{i=0}^{k}\alpha_i^{d_i}(1-\alpha_i)^{n_i-d_i}\right) \\ &= \sum_{i=0}^{k}\log(\alpha_i^{d_i}) + \log((1-\alpha_i)^{n_i-d_i}) \\ &= \sum_{i=0}^{k} d_i\log(\alpha_i) + (n_i-d_i)\log(1-\alpha_i) \end{aligned}$$

entonces

$$\begin{aligned}
\frac{\partial \log(L(S))}{\partial \alpha_i} &= \frac{d_i}{\alpha_i} - \frac{n_i - d_i}{1-\alpha_i} \\
&\Rightarrow \ \frac{d_i}{\alpha_i} - \frac{n_i - d_i}{1-\alpha_i} = 0 \\
&\Rightarrow \ \frac{d_i - d_i\alpha_i - n_i\alpha_i + d_i\alpha_i}{\alpha_i(1-\alpha_i)} = 0 \\
&\Rightarrow \ \frac{d_i - n_i\alpha_i}{\alpha_i(1-\alpha_i)} = 0 \\
&\Rightarrow \ d_i - n_i\alpha_i = 0 \\
&\Rightarrow \ \widehat{\alpha_i} = \frac{d_i}{n_i}
\end{aligned}$$

Además

$$\frac{\partial^2 \log(L(S))}{\partial \alpha_i^2} = -\frac{d_i}{\alpha_i^2} - \frac{n_i - d_i}{(1-\alpha_i)^2} \quad y \quad \frac{\partial^2 \log(L(S))}{\partial \alpha_i \partial \alpha_j} = 0$$

por lo cual obtenemos la matriz siguiente

$$H = \begin{pmatrix}
-\frac{d_0}{\alpha_0} - \frac{n_0 - d_0}{(1-\alpha_0)^2} & 0 & \cdots & 0 \\
0 & -\frac{d_1}{\alpha_1} - \frac{n_1 - d_1}{(1-\alpha_1)^2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & -\frac{d_k}{\alpha_k} - \frac{n_k - d_k}{(1-\alpha_k)^2}
\end{pmatrix}$$

ahora veamos si es negativa definida.

La matriz $(-1) \times H$ es definida positiva puesto que

$$(-1) * H = \begin{pmatrix}
\frac{d_0}{\alpha_0} + \frac{n_0 - d_0}{(1-\alpha_0)^2} & 0 & \cdots & 0 \\
0 & \frac{d_1}{\alpha_1} + \frac{n_1 - d_1}{(1-\alpha_1)^2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{d_k}{\alpha_k} + \frac{n_k - d_k}{(1-\alpha_k)^2}
\end{pmatrix}$$

por lo cual la matriz H es definida negativa, de igual manera podemos decir que α_i $i = 0, 1, \ldots, k$ es un máximo. Asimismo el estimador máximo verosímil es

$$\widehat{S}(t) = \prod_{i|t_i<t} \left(1 - \frac{d_i}{n_i}\right)$$

3.3.4. Varianza del estimador Kaplan-Meier

Para calcular la varianza del estimador Kaplan-Meier primero recordemos el método delta, el cual utilizaremos como base

Método Delta: Sea $\mathbb{X}$ una variable aleatoria con media α y varianza σ^2, y sea g una función infinitamente diferenciable, entonces

$$\mathbb{E}[g(\mathbb{X})] \approx g(\alpha)$$

$$\mathbb{V}ar[g(\mathbb{X})] \approx g'(\alpha)^2 \mathbb{V}ar(\mathbb{X}) = g'(\alpha)^2 \sigma^2$$

Sea t tal que $t_k \leq t < t_{k+1}$ tenemos que

$$\log(\widehat{S}(t)) = \sum_{j=1}^{k} \log\left(1 - \frac{d_j}{n_j}\right) = \sum_{j=1}^{k} \log(1 - \widehat{q_j}) = \sum_{j=1}^{k} \log(\widehat{p_j})$$

entonces

$$\mathbb{V}ar\left[\log(\widehat{S}(t))\right] = \mathbb{V}ar\left(\sum_{j=1}^{k} \log(\widehat{p_j})\right) = \sum_{j=1}^{k} \mathbb{V}ar(\log(\widehat{p_j}))$$

Si observamos el último término de la igualdad se supone que los $\widehat{p_j}$ son independientes, además que los $\widehat{p_j}$ tienen dsitrbucion Bernoulli con media $\widehat{p_j}$ y varianza $\frac{\widehat{p_j}(1-\widehat{p_j})}{n_j}$

Ahora si aplicamos el método Delta, tenemos

$$\mathbb{V}ar(\log(\widehat{p_j})) \approx \left(\frac{1}{\widehat{p_j}}\right)^2 \frac{\widehat{p_j}(1-\widehat{p_j})}{n_j} = \frac{1-\widehat{p_j}}{n_j \widehat{p_j}}$$

puesto que $\widehat{p_j}$ se distribuye Bernoulli con parámetro p_j. Además se concluye que

$$\mathbb{V}ar(\log[\widehat{S}(t)]) \approx \sum_{j=1}^{k} \frac{1-\widehat{p_j}}{n_j \widehat{p_j}}$$

Pero estamos buscando la varianza del $\widehat{S}(t)$ no de su logaritmo, por lo cual debemos aplicar de nuevo el método Delta ahora para $\log(\widehat{S}(t))$

$$\mathbb{V}ar(\log[\widehat{S}(t)]) \approx \left(\frac{1}{\widehat{S}(t)}\right)^2 \mathbb{V}ar(\widehat{S}(t))$$

entonces

$$\mathbb{V}ar(\widehat{S}(t)) \approx \widehat{S}(t)^2 \, \mathbb{V}ar(\log[\widehat{S}(t)])$$

Por lo tanto

$$\mathbb{V}ar(\widehat{S}(t)) \approx \widehat{S}(t)^2 \sum_{j=1}^{k} \frac{1-\widehat{p_j}}{n_j \widehat{p_j}}$$

Recordemos que $p_j = 1 - \frac{d_j}{n_j}$ por lo cual

$$\mathbb{V}ar(\widehat{S}(t)) \approx \widehat{S}(t)^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}$$

Y para cualquier t

$$\mathbb{V}ar(\widehat{S}(t)) \approx \widehat{S}(t)^2 \sum_{j|t_j<t}^{k} \frac{d_j}{n_j(n_j - d_j)}$$

3.3.5. Intervalos de confianza

Se construye intervalos para cada valor de t donde se evalue el estimador Kaplan-Meier, y la forma general de estos intervalos son

$$S(t) \in \left(\widehat{S}(t) \pm Z_{1-\alpha/2}\sqrt{\mathbb{V}ar[\widehat{S}(t)]}\right)$$

puesto que se tiene el supuesto que sigue una distribución asintótica con media normal $\widehat{S}(t)$ y varianza $\mathbb{V}ar[\widehat{S}(t)]$.

Aunque estos intervalos tienen ciertos inconvenientes por lo cual se recomienda utilizar cierta transformación la cual es mediante el logaritmo. Este nuevo intervalo es obtenido mediante el supuesto que el logaritmo del estimador Kaplan-Meier sigue una dis-

tribución asintótica normal con media $\log(S(t))$ y varianza $\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}$.

Por tanto, un intervalo de un $100(1-\alpha)\,\%$ de confianza para $\log(S(t))$ en un tiempo definido t es

$$log(S(t)) \in \left(\log(\widehat{S}(t)) \pm Z_{1-\alpha/2} \sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}} \right)$$

Asimismo el intervalo para $S(t)$ es

$$S(t) \in \left(\widehat{S}(t) * \exp\left[-Z_{1-\alpha/2} \sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}} \right], \widehat{S}(t) * \exp\left[Z_{1-\alpha/2} \sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}} \right] \right)$$

3.3.6. Comparación de funciones de supervivencia

Usualmente en los estudios de supervivencia uno de los problemas más importantes es comparar poblaciones, es decir, si alguna característica (género, aplicación de algún tratamiento, nivel educativo, etcétera) genera una diferencia de supervivencia en la población. Una manera común de realizar las comparaciones entre poblaciones es mediante el planteamiento de una prueba paramétrica, la cual requiere el cumplimiento de varios supuestos para su implementación. En los estudios de supervivencia, dado la naturaleza de sus datos (existen datos censurados), es más difícil construir pruebas paramétricas para hacer estas comparaciones, por lo cual una alternativa viable es realizar una prueba no paramétrica.

3.3.6.1. Prueba Log-rank, Mantel-Haenszel

Comencemos con el caso sencillo, cuando tenemos sólo dos grupos. Supongamos que se tienen k tiempos de falla en la muestra, donde se combinan los dos grupos: $t_1, t_2, \ldots, t_k$ y que en el tiempo t_i, hay d_{1i} sujetos del primer grupo y d_{2i} del segundo que fallan, con $i = 1, 2, \ldots, k$. Además supongamos también que existen n_{1i} y n_{2i} individuos en riesgo justo antes del tiempo t_i del primer y segundo grupo, respectivamente. Por lo cual hay $d_i = d_{1i} + d_{2i}$ fallas, de $n_i = n_{1i} + n_{2i}$ sujetos en riesgo en t_i.

Se quiere probar la hipótesis de igualdad en las funciones de supervivencia entre poblaciones, es decir

$$H_0 : S_1(t) = S_2(t) \ \ \forall t \ \ vs. \ \ H_a : S_1(t) \neq S_2(t) \ \ p.a. \ t > 0$$

Así que una forma de realizar esta prueba es tomar en cuenta la diferencia entre el número de fallas de los grupos, que ocurren en cada tiempo de falla y el número esperado bajo la hipótesis nula.

Podemos representar dicha información mediante una tabla de contingencia, lo cual nos permitirá tener un mejor manejo de la misma. Esta tabla tiene por renglones a los grupos y el estado de supervivencia en las columnas.

Tabla 3.2: Tabla usada para la prueba de igualdad de supervivencia entre dos grupos en el tiempo t_i

Grupo	Fallas	Sin falla	En riesgo antes de t_i
I	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
II	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Si de esta tabla tomamos fijos los marginales totales ($d_i, n_i - d_i, n_i$), y la hipótesis nula es verdadera, las entradas de la tabla quedan determinadas si conocemos únicamente d_{1i}. Por lo cual D_{1i} es una variable aleatoria con distribución hipergeométrica, es decir

$$D_{1i} \backsim \text{hipergeométrica}(n_i, d_i, n_{1i})$$

Entonces

$$\mathbb{P}(D_{1i} = d_{1i}) = \frac{\binom{d_i}{d_{1i}}\binom{n_i - d_i}{n_{1i} - d_{1i}}}{\binom{n_i}{n_{1i}}}$$

Puesto que la tabla queda definida con solo conocer d_{1i}, utilizaremos esta variable para realizar la prueba. Dado que d_{1i} tiene distribución hipergeométrica se tiene

$$e_{1i} = \mathbb{E}(D_{1i}) = n_{1i}\frac{d_i}{n_i}$$

$$V_{1i} = \mathbb{V}ar(D_{1i}) = \frac{n_{1i}n_{2i}d_i(n_i - d_i)}{n_i^2(n_i - 1)}$$

Entonces, se debe tomar en cuenta esta comparación para cada t_i, con $i = 1, 2, \ldots, k$. Por lo cual el estadístico de prueba es

$$Q = \sum_{i=1}^{k}(d_{1i} - e_{1i})$$

Además este estadístico tiene las siguientes características

- $\mathbb{E}(Q) = 0$
- $\mathbb{V}ar(Q) = \sum_{i=1}^{k} V_{1i}$

Por lo cual, si las supervivencias en las dos poblaciones son iguales

$$Q = \frac{\sum_{i=1}^{k}(d_{1i} - e_{1i})}{\sqrt{\sum_{i=1}^{k} V_{1i}}} \backsim N(0, 1)$$

De igual manera

$$Q^2 \backsim \chi^2_{(1)}$$

Por lo tanto, la regla de decisión a un nivel de significancia α es:

- $Q^2 > \chi^2_{(1),\,1-\alpha}$ se rechaza H_0
- $Q^2 < \chi^2_{(1),\,1-\alpha}$ no se rechaza H_0

3.3.6.2. Comparación de m poblaciones

En este caso se quiere realizar la prueba

$$H_0 : S_1(t) = S_2(t) = \cdots = S_m(t) \;\; \forall t \;\; vs. \;\; H_\alpha : S_r(t) \neq S_s(t) \;\; p.a.\; r \neq s \;\; r, s = 1, 2, \ldots, m$$

Para esta prueba denotemos a $\bar{d}_j$ como el vector de fallas al tiempo t_j, $j = 1, 2, \ldots, m$, con vector de medias $\mathbb{E}(\bar{d}_j)$ y matriz de varianzas y covarianzas, cuya entrada (i, k) está

definida por

$$\mathbb{V}ar(d_{ij}) = \frac{n_{ji}n_{ji+1}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

$$Covar(d_{ji}, d_{jk}) = \frac{n_{ji}n_{jk}d_j(n_j - d_j)}{n_j^2(n_j - 1)},\ i \neq k$$

Por lo cual si se suman todos los tiempos de falla se obtienen

$$B = \sum_{j=1}^{k}(\bar{d_j} - \mathbb{E}(\bar{d_j}))$$

$$V = \sum_{j=1}^{k}\mathbb{V}ar(\bar{d_j})$$

Para esta prueba Mantel-Hanszel proponen realizarla sobre k funciones de supervivencia iguales, a través de la forma cuadrática

$$Q = B^t V^- B$$

que bajo el supuesto que se cumpla la hipótesis nula, se distribuye asintóticamente como $\chi^2_{(m-1)}$. Con V^- la inversa generalizada de V.

Por lo tanto, la regla de decisión a un nivel de significancia α es:

- $Q > \chi^2_{m-1,\,1-\alpha}$ se rechaza H_0
- $Q < \chi^2_{m-1,\,1-\alpha}$ no se rechaza H_0

3.3.6.3. Prueba de Wilcoxon

La prueba de Wilcoxon es similar a la prueba log-rank, pero Wilcoxon pondera las diferencias de la log-rank a través del número de sujetos en riegos en el tiempo t_i, por lo cual el estadístico queda de la manera siguiente

$$Q_W = \sum_{i=1}^{k} n_i(d_{1i} - e_{1i})$$

Además se mantienen las propiedades del estadístico de la log-rank con una ligera modificación en la varianza

$$V_W = \sum_{i=1}^{k} n_i^2 V_{1i}$$

Así que, si las funciones de supervivencia son iguales

$$W = \frac{Q_W}{V_W} \backsim \chi^2_{(m-1)}$$

3.4. Riesgos proporcionales de Cox

Supongamos que se quiere saber qué tanto influye una variable o característica en el riego de sufrir la falla para un individuo. Para ello existe un modelo para la función de riesgo ($h(t)$), el cual es el modelo de riesgos proporcionales de Cox.

En primera instancia veamos qué características tiene dicho modelo:

- Se tiene una población base $\bar{Z} = \bar{0}$, con riesgo $h_0(t)$.
- Se supone que existe $\rho > 0$ tal que $h(t|\bar{Z}) = \rho h_0(t)$, por lo que si $\rho > 1$ tiene mayor riesgo de presentar la falla la población con covariables $\bar{Z}$ y si $0 < \rho < 1$ la población base presentará mayor riesgo de sufrir la falla.

  Además ρ no depende de t, puesto que

$$\rho = \frac{h(t|\bar{Z})}{h_0(t)} \quad \forall t$$

Ahora veamos que propiedades tiene $\rho(\bar{Z})$

- $\rho(\bar{Z}) > 0$
- $\rho(\bar{0}) = 1$
- Su forma paramétrica es $\rho(\bar{Z}) = \exp(\bar{\beta}'\bar{Z})$ con $\bar{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$

Por lo tanto el modelo de riesgos proporcionales de Cox es de la forma

$$h(t|\bar{Z}) = \exp(\bar{\beta}'\bar{Z})h_0(t)$$
$$\frac{h(t|\bar{Z})}{h_0(t)} = \exp(\bar{\beta}'\bar{Z})$$

Así que si queremos comparar poblaciones cuyas covariables sean $\bar{Z}_1$ y $\bar{Z}_2$ tenemos

$$\frac{h(t|\bar{Z}_1)}{h(t|\bar{Z}_2)} = \exp(\bar{\beta}'(\bar{Z}_1 - \bar{Z}_2))$$

De la expresión anterior podemos destacar que es un modelo tipo log-lineal, puesto que se crea una relación lineal al aplicar logaritmo al cociente

$$\log\left(\frac{h(t|\bar{Z})}{h_0(t)}\right) = \bar{\beta}'\bar{Z} = \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p$$

Ahora veamos cuál es la interpretación de estos parámetros. Consideremos el modelo con un solo regresor continuo Z y tomemos dos valores consecutivos de esta covariable z y $z+1$, entonces tenemos:

$$Z = z+1 \Rightarrow \log\left(\frac{h(t|z+1)}{h_0(t)}\right) = \beta_1(z+1) = \beta_1 z + \beta_1$$

$$Z = z \Rightarrow \log\left(\frac{h(t|z)}{h_0(t)}\right) = \beta_1 z$$

Así que

$$\log\left(\frac{h(t|z+1))}{h_0(t)}\right) - \log\left(\frac{h(t|z))}{h_0(t)}\right) = \log\left(\frac{h(t|z+1))}{h(t|z)}\right) = \beta_1 z + \beta_1 - \beta_1 z = \beta_1$$

Por lo cual se puede interpretar a β_1 como el cambio en el logaritmo del cociente de riesgos por unidad de cambio en la covariable. Asimismo si tomamos ahora $\exp(\beta_1)$ podemos interpretarlo como el cambio en el cociente de riesgos por el incremento en una unidad de la covariable. Como se vio con anterioridad, es común que se estimen los valores o las funciones de riesgo y supervivencia, en el caso que se haya estimado a β_1 se interpreta como el cambio promedio o esperado por unidad de cambio en la covariable.

En el caso de las variables categóricas, habrá tantos parámetros como categorías menos una, además $\exp(\beta_j)$ sera el cambio en el cociente de riesgos entre un sujeto en la categoría j de dicha variable, contra uno de la categoría basal.

Por ejemplo, sea el modelo con un solo regresor categórico, con n categorías , entonces tenemos

$$h(t|\bar{Z}) = h_0(t) \exp\left(\sum_{j=0}^{n-1} \beta_j Z_j\right)$$

donde $h_0(t)$ es el riesgo de sufrir la falla para los individuos que pertenecen a la categoría base, $Z_j = 1$ si el individuo pertenece a la categoría j y $Z_j = 0$ si el individuo no pertenece a la categoría j.

Así que

$$\log\left(\frac{h(t|Z_j = 1))}{h_0(t)}\right) = \beta_j$$

Entonces β_j se puede interpretar como el cambio en el logaritmo del cociente de riesgos, al comparar un individuo que pertenece a la categoría j con uno de la categoría base.

3.4.1. Estimación en el modelo de riesgos proporcionales y pruebas de hipótesis

En el modelo de riesgos proporcionales para estimar los parámetros $\bar{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, se utiliza el método de máxima verosimilitud, para construir nuestra verosimilitud contamos con la siguiente información $(T_i, \delta_i, \bar{Z}_i)$ con $i = 1, 2, 3, \ldots, n$. Supongamos que tenemos k tiempos de falla distintos y sin empates por ahora, asimismo sea R_i el i-ésimo conjunto de riesgo, es decir el conjunto de los n_i individuos en riesgo al tiempo t_i

Para construir la verosimilitud tomemos en cuenta la probabilidad de que un sujeto

muere en algún tiempo t_j, por lo cual este sujeto tiene covariables $\bar{Z}_j$.

$$\begin{aligned}&\mathbb{P}[\text{Un individuo con covariables } \bar{Z}_j \text{ sufra la falla en } t_j | \text{Pertenece al conjunto de riesgo } R_j] \\ &= \frac{\mathbb{P}[\text{Un individuo con covariables } \bar{Z}_j \text{ sufra la falla en } t_j]}{\mathbb{P}[\text{Conjunto de riesgo } R_j]} \\ &= \frac{h(t_j|\bar{Z}_j)}{\sum \text{Riesgo de todos los sujetos del conjunto } R_j} = \frac{h(t_j|\bar{Z}_j)}{\sum_{l\in R_j} h(t_l|\bar{Z}_l)}\end{aligned}$$

que bajo el supuesto de riesgo proporcionales tenemos

$$\frac{h(t_j|\bar{Z}_j)}{\sum_{l\in R_j} h(t_l|\bar{Z}_l)} = \frac{h_0(t_j)\exp(\bar{\beta}'\bar{Z}_j)}{\sum_{l\in R_j} h_0(t_l)\exp(\bar{\beta}'\bar{Z}_l)} = \frac{\exp(\bar{\beta}'\bar{Z}_j)}{\sum_{l\in R_j}\exp(\bar{\beta}'\bar{Z}_l)}$$

Por lo cual podemos decir que la verosimilitud es el producto de la expresión anterior sobre todos los tiempos de falla.

$$\prod_{j=1}^{k} \frac{\exp(\bar{\beta}'\bar{Z}_j)}{\sum_{l\in R_j}\exp(\bar{\beta}'\bar{Z}_l)}$$

Podemos apreciar que esta verosimilitud no es de la forma usual, ya que el número de términos no es el total de individuos sino el de fallas, además los elementos censurados no aportan información al numerador, pero sí influyen en el denominador puesto que están en el conjunto de riesgo, asimismo se puede observar que si quitamos o aumentamos un individuo observado al conjunto de riesgo la verosimilitud cambia.

Por lo tanto podemos reescribir la verosimilitud como

$$L(\bar{\beta}, \bar{t}, \delta_j, \bar{Z}_j) = \prod_{j=1}^{n} \left(\frac{\exp(\bar{\beta}'\bar{Z}_j)}{\sum_{l\in R_j}\exp(\bar{\beta}'\bar{Z}_l)} \right)^{\delta_j}$$

$$donde\ \delta_j = 1\ si\ es\ falla\ y\ \delta_j = 0\ si\ es\ cesnura.$$

Asimismo

$$\log(L(\bar{\beta}, \bar{t}, \delta_j, \bar{Z}_j)) = \sum_{j=1}^{n} \delta_j \left[\bar{\beta}\bar{Z}_j - \log\left(\sum_{l\in R_j} \exp\{\bar{\beta}'\bar{Z}_l\} \right) \right]$$

Ahora apliquemos los procesos usuales a la expresión anterior para obtener los esti-

madores de los parámetros que buscamos, por lo cual si derivamos parcialmente con respecto a β_k

$$U_k(\bar{\beta}) = \frac{\partial \log(L(\bar{\beta}))}{\partial \beta_k} = \sum_{j=1}^{n} \delta_j \left[z_{kj} - \frac{\sum_{l \in R_j} z_{kj} \exp(\bar{\beta}' \bar{Z}_l)}{\sum_{l \in R_j} \exp(\bar{\beta}' \bar{Z}_l)} \right]$$

Los estimadores β_k se obtienen igualando estas expresiones a cero, mismas que se resuelven mediante métodos numéricos.

En cuanto a las pruebas, éstas se realizan para verificar la significancia global del modelo y para verificar las significancia de cada covariable. Esta última es de suma importancia puesto que sirve para saber o determinar el impacto que tiene una característica de los sujetos en el riegos de presentar la falla. En términos estadísticos las pruebas serían

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p \;\; vs. \;\; H_a : \beta_j \neq 0 \;\; p.a. \; j = 1, 2, \ldots, p$$
$$H_0 : \beta_j = 0 \;\; vs. \;\; H_a : \beta_j \neq 0 \;\; j = 1, 2, \ldots, p$$

respectivamente.

3.4.2. Bondad de ajuste y análisis de residuos

En cuanto se ha estimado el modelo es sumamente importante saber qué tan bien representa o se ajusta a nuestros datos, además debemos de recordar que el modelo se construyó con base en el supuesto de riesgos proporcionales, así que es necesario verificar este supuesto. Como vimos con anterioridad el modelo que se obtiene a partir del supuesto de riesgos proporcionales es tipo log-lineal, por lo cual es de suponerse que para verificar estos supuestos no basemos en los residuos al igual que como se hace en el modelo de regresión lineal, pero existen una pequeña diferencia en los modelos de supervivencia existe datos censurados, lo cual nos da varios tipos de residuos.

3.4.2.1. Contraste de la razón de verosimilitud

Para realizar la prueba $H_0 : \beta = \bar{0}$ versus $H_a : \beta \neq \bar{0}$, primero recordemos la función de verosimilitud parcial

$$L(\bar{\beta}, \bar{t}, \delta_j, \bar{Z}_j) = \prod_{j=1}^{n} \left(\frac{\exp(\bar{\beta}' \bar{Z}_j)}{\sum_{l \in R_j} \exp(\bar{\beta}' \bar{Z}_l)} \right)^{\delta_j}$$

$$donde\ \delta_j = 1\ si\ es\ falla\ y\ \delta_j = 0\ si\ es\ cesnura.$$

de la cual obtenemos una estimación de los coeficientes $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ y cuyo logaritmo es

$$\log(L(\bar{\beta}, \bar{t}, \delta_j, \bar{Z}_j)) = \sum_{j=1}^{n} \delta_j \left[\bar{\beta} \bar{Z}_j - \log \left(\sum_{l \in R_j} \exp\{\bar{\beta}' \bar{Z}_l\} \right) \right]$$

Así que para el contraste de razón de verosimilitud se utiliza la función de verosimilitud parcial evaluada en $\widehat{\beta}$, $L(\widehat{\beta})$, y evaluada en $\bar{0}$, $L(\widehat{0})$.

El estadístico se define como:

$$X = 2(\log L(\widehat{\beta}) - \log L(\bar{0}))$$

que bajo la hipótesis nula sigue una distribución χ^2 con p grados de libertad.

Por lo tanto, la regla de decisión a un nivel de significancia α es:

- $X > \chi^2_{p,\, 1-\alpha}$ se rechaza H_0
- $X^2 < \chi^2_{p,\, 1-\alpha}$ no se rechaza H_0

3.4.2.2. Residuos escalados de Schoenfeld

Son semejantes a los residuos de mínimos cuadrados. Se definen como

$$r^*_{ij} = k \mathbb{V}ar(\bar{\beta}) r^s_{ij}$$

donde r_{ij}^s son los residuos de Schoenfeld[2], k es el número de fallas y $\mathbb{V}ar(\bar{\beta})$ la matriz p x p de varianzas y covarianzas del vector de parámetros estimados al ajustar el modelo de riesgos proporcionales. Grambsch y Therneau (1994) demostraron que si $\widehat{\beta_j}$ es el coeficiente estimado al ajustar el modelo de riesgos proporcionales sin covariables dependientes del tiempo, entonces

$$\mathbb{E}(r_{ij}^*) \approx \beta_j(t_i) - \widehat{\beta_j}$$

Por lo cual al graficar $r_{ij}^* + \widehat{\beta_j}$ contra el tiempo o una función que dependa de él, sirve para observar la naturaleza y proporcionalidad de los riesgos. Al ajustar una recta a la gráfica ésta debe verse aproximadamente horizontal si los riesgos son proporcionales.

3.4.2.3. Residuos de martingalas

Se define el residuo de martingala para el i-ésimo sujeto de la siguiente manera

$$r_i = \delta_i - \widehat{\Lambda}(t_i) \;\; i = 1, \cdots, n$$

donde $\widehat{\Lambda}$ es la función de riesgo acumulado estimada. De dicha expresión podemos deducir que los residuos de martingalas son la diferencia entre el número observado de fallas por cada sujeto i, desde el inicio hasta el tiempo t_i, y el número esperado de falla basado en el modelo.

Los residuos de martingalas tiene media cero y son aproximadamente no correlacionados en muestras grandes. Este tipo de residuos sirven para verificar que las variables ingresan al modelo de manera lineal, es decir, si una variable tiene una relación lineal con el riesgo, los datos (puntos) en la gráfica donde se comparan los residuos de martingalas contra los valores de dicha variable deben verse distribuidos de manera aleatoria.

3.4.2.4. Residuos de devianza

Los residuos de martingalas presentan un inconveniente, ya que tienden a ser asimétricos con sesgo hacia la derecha, por lo cual se recomienda utilizar los residuos de

[2] $r_{ij}^s = Z_{ij}(t_i) - Z_j(t_i),\ j = 1, \ldots, p\ i = 1, \ldots, k$,

devianza, lo cuales son una transformación de los residuos de martingalas y se definen para cada observación i como

$$\widehat{D_i} = signo(r_i)\sqrt{-2[r_i + \delta_i \log(\delta_i - r_i)]}$$

Los residuos de devianza tiene media cero y varianza uno, además son negativos para observaciones cuyo tiempo de supervivencia es menor del esperado. Al igual que los residuos de martingalas pueden ser graficados contra los valores de cada variable, además sirven para detectar observaciones con residuos grandes, esto se observa si el residuo es mayor que 2 o 3 en valor absoluto.

3.5. Modelo extendido de Cox

Puesto que existen ocasiones en las cuales el modelo de Cox estándar no representa con veracidad a los datos, tal es el caso cuando existen variables que dependen del tiempo, o, en ciertas ocasiones algunas variables no cumplen el supuesto de riesgos proporcionales, por lo cual existe una extensión del modelo de Cox estándar para dar solución a estos problemas

En primera instancia debemos conocer las características de las nuevas variables, nos enfocaremos en las que dependen del tiempo. Las variables dependientes del tiempo pueden ser: intrínsecamente dependientes del tiempo o bien definidas para la evaluación de la hipótesis de riesgos proporcionales a partir de predictores independientes del tiempo. Un ejemplo de una variable que es independiente del tiempo, pero se requiere la interacción de dicha variable con el tiempo, es el siguiente: supongamos que la variable W toma lo valores (0,1), W es independiente del tiempo, pero $W \times g(t)$, donde $g(t)$ es una función que depende del tiempo, no lo es. Si W toma los valores 0 y 1 entonces $W \times g(t) = g(t)\ si\ W = 1\ y\ W \times g(t) = 0\ si\ W = 0$. Las funciones más comunes para realizar una interacción de una variable independiente con el tiempo son

$$g(t) = \begin{cases} t \\ \log(t) \\ \exp(t) \\ \sqrt{t} \end{cases}$$

Por otro lado, un ejemplo de una función que se utiliza en el contraste de la hipótesis de riesgos proporcionales es la denominada función de cambio de punto, la cual se define como

$$g(t) = \begin{cases} 1 & si \quad t \geq t_0 \\ 0 & si \quad t < t_0 \end{cases}$$

esta función se utiliza en ocasiones para hacer que una variable que no cumple el supuesto de riesgos proporcionales, lo cumpla. Supongamos que la variable E no cumple el supuesto de riesgos proporcionales y a partir del tiempo t_0 se observa que existe proporcionalidad en sus riesgos, por lo cual se define la variable $E * g(t)$, donde $g(t)$ es una función de cambio de punto, para que la variable E cumpla el supuesto de riesgos proporcionales y pueda seguir influyendo en el estudio.

Otro tipo de variables dependientes del tiempo son aquellas que intrínseca o internamente dependen del tiempo, ya que sus valores varían por la propias características del sujeto. Por ejemplo, el nivel de obesidad, estado de empleo, estado de fumador.

Existe otro tipo de variables dependientes del tiempo, ya que sus valores varían a lo largo de éste, pero no por causas intrínsecas al individuo sino debido a características externas. A estas variables se les denomina auxiliares o externas. Por ejemplo, el índice de contaminación o la temperatura.

3.5.1. Verosimilitud del modelo de Cox extendido

Puesto que tenemos ahora variables con diferentes características (dependen del tiempo), la verosimilitud cambia, en primer lugar recordemos la verosimilitud para el caso donde las variables no dependen del tiempo, la cual es

$$L(\bar{\beta}, \bar{t}, \delta_j, \bar{Z}_j) = \prod_{j=1}^{n} \left(\frac{\exp(\bar{\beta}' \bar{Z}_j)}{\sum_{l \in R_j} \exp(\bar{\beta}' \bar{Z}_l)} \right)^{\delta_j}$$

Esta verosimilitud compara el riesgo de un individuo con covariables $\bar{Z}_j$, contra sujetos con covariables $\bar{Z}_l$ en el conjunto de riesgo R_j al tiempo t_j. Pero no tenemos alguna razón para suponer que las covariables $\bar{Z}_i$ del i-esimo sujeto sean las mismas para todos los tiempos t_j. Así que la verosimilitud para el modelo de riesgos proporcionales con

covariables que dependen del tiempo es

$$L(\bar{\beta}, \bar{t}, \delta_j, \bar{Z}_j) = \prod_{j=1}^{n} \left(\frac{\exp(\bar{\beta}' \bar{Z}_j(t_j))}{\sum_{l \in R_j} \exp(\bar{\beta}' \bar{Z}_l(t_j))} \right)^{\delta_j}$$

Se puede apreciar que ahora esta verosimilitud compara el riesgo de un sujeto con covariables $Z_j(t_j)$ al tiempo t_j en el conjunto de riesgo R_j.

3.5.2. Formulación del modelo de Cox extendido para variables dependientes del tiempo

El modelo de Cox extendido se define como

$$h(t|\bar{Z}) = h_0(t) \exp\left\{ \bar{\beta_1}' \bar{Z}_1 + \bar{\beta_2}' \bar{Z}_2 \right\}$$

donde $\bar{Z}_1$ es la matriz de variables que no dependen del tiempo y $\bar{Z}_2$ las que sí dependen del tiempo.

Supongamos que tenemos p_1 predictores independientes del tiempo y p_2 predictores dependientes del tiempo, entonces el modelo de Cox extendido sería

$$h(t|\bar{Z}) = h_0(t) \exp\left\{ \sum_{j=1}^{p_1} \beta'_{1j} Z_j + \sum_{j=1}^{p_2} \beta'_{2j} Z_j(t) \right\}$$

Los procesos de estimación son similares a los del modelo de riesgos proporcionales estándar, la diferencia radica en los datos, puesto que ahora el tiempo está dado por intervalos, el tiempo ahora es de la forma $[0, t_{1j}), [t_{1j}, t_{2j}), \ldots, [t_{ij}, t_{(i+1)j}), \ldots$, donde t_{ij} es el tiempo en el cual la variable dependiente del tiempo, cambió en la ocasión anterior y $t_{(i+1)j}$ es el tiempo donde la variable vuelve a cambiar. A este tipo de formato se le conoce como ”proceso de conteo” o ” (inicio, pausa)”.

3.5.3. Formulación de los riesgos

Dados dos conjuntos de predictores $\bar{Z}^*(t) = ((Z_1^*, \ldots, Z_{p_1}^*), Z_1^*(t), \ldots, Z_{p_2}^*(t))$ y $\bar{Z}(t) = ((Z_1, \ldots, Z_{p_1}), Z_1(t), \ldots, Z_{p_2}(t))$ entonces tenemos que

$$\begin{aligned}\frac{h(t|\bar{Z}^*(t))}{h(t|\bar{Z}(t))} &= \frac{h_0(t)\exp\left\{\sum_{j=1}^{p_1}\beta'_{1j}Z_j^* + \sum_{j=1}^{p_2}\beta'_{2j}Z_j^*(t)\right\}}{h_0(t)\exp\left\{\sum_{j=1}^{p_1}\beta'_{1j}Z_j + \sum_{j=1}^{p_2}\beta'_{2j}Z_j(t)\right\}} \\ &= \exp\left\{\sum_{j=1}^{p_1}\beta'_{1j}(Z_j^* - Z_j) + \sum_{j=1}^{p_2}\beta'_{2j}(Z_j^*(t) - Z_j(t))\right\}\end{aligned}$$

Un ejemplo suponiendo una única variable E que tome los valores (0,1) y utilizando $g(t) = t$ tenemos el modelo de Cox extendido

$$h(t|\bar{Z}) = h_0(t)\exp\{\beta'_{11}E + \beta'_{21}(E \times t)\}$$

Entonces el cociente de riesgos es

$$\frac{h(t, E=1)}{h(t, E=0)} = \exp(\beta'_{11} + \beta'_{21}t)$$

Observemos que no se cumple la hipótesis de proporcionalidad de riesgos. Por otro lado, cabe notar que cada β_{2j} para $j = 1, 2, \ldots, p_2$ es independiente del tiempo y representa el efecto global de la variable a lo largo del tiempo, es decir, teniendo en cuenta todas las observaciones de la variable a lo largo del tiempo.

Capítulo 4

Análisis de supervivencia en obesidad

En este apartado se mostrarán los resultados obtenidos. La base de datos que se analizó consta de 1145 registros que corresponden a 269 personas, esto se debe a que los regitros corresponden a cada visita de las 269 personas al Hospital Infatil para anotar los cambios en su dieta, edad, peso y talla. Las variables que componen a la base de datos son las siguientes

Variable	Descripción	Tipo de variable
folio	código que identifica al paciente	
visita	número de visita	
cat3	censura=0 falla=1	
zbfa	índice de masa corporal estandarizado	continua - dependiente del tiempo
edad	edad en años del paciente a la fecha de visita	continua - dependiente del tiempo
sexo	género del paciente (1=Hombre, 0=Mujer)	discreta - independiente del tiempo

cat_kcal	ingesta diaria de calorías (1=Alto 0=Normal)	discreta - dependiente del tiempo
carbo_tot_dia	ingesta diaria promedio de carbohidratos	continua - dependiente del tiempo
lipid_tot_dia	ingesta diaria promedio de lípidos	continua - dependiente del tiempo
prote_tot_dia	ingesta diaria promedio de proteínas	continua - dependiente del tiempo
cat_peso	Peso al nacer (0=Bajo peso 1=Peso normal 2=Peso alto)	discreta - independiente del tiempo
actfisica_esc	Realiza actividad física(Sí=1 No=0)	discreta - independiente del tiempo

Recordemos que la falla o evento de interés, es cuando un paciente disminuye el cinco por ciento de su índice de masa corporal estandarizado inicial.

4.1. Estimador Kaplan-Meier

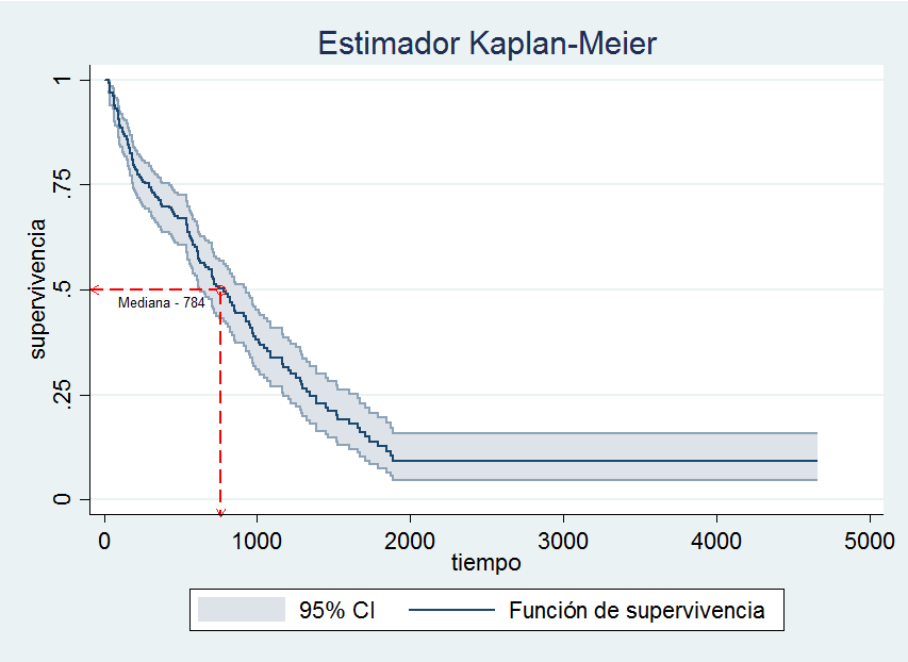
Debemos observar el estimador Kaplan-Meier en general y si existe diferencia en la función de supervivencia dependiendo de variables que se atribuyen son influyentes en que el individuo presente la falla o no. En la tabla siguiente se puede observar que el tiempo medio que necesita un individuo para disminuir el 5 % de su índice de masa corporal estandarizado inicial es de 784 días, es decir, más de 2 años.

Tabla 4.2: Resumen

Sujetos	Eventos	Mediana	Intervalo	
269	160	784	626	924

De igual manera se muestra la media en la gráfica siguiente

Figura 4.1: Estimador Kaplan-Meier



4.1.1. Por sexo

En la tabla siguiente se presenta la media de supervivencia para hombres y mujeres de la cual podemos decir que fueron más los hombres que lograron disminuir el cinco por ciento de su índice de masa corporal estandarizado inicial en comparación con las mujeres, aunque el tiempo medio que requieren es mayor puesto que las mujeres necesitan 702 días y los hombres 807, lo cual nos haría suponer que las mujeres tienden a disminuir su índice de masa corporal estandarizado más rápido.

Tabla 4.3: Media del tiempo de supervivencia por sexo

Sexo	Sujetos	Eventos	Mediana	Error estándar	Intervalo 95 %	
Mujer	119	68	702	102.335	548	989
Hombre	150	92	807	86.264	621	954

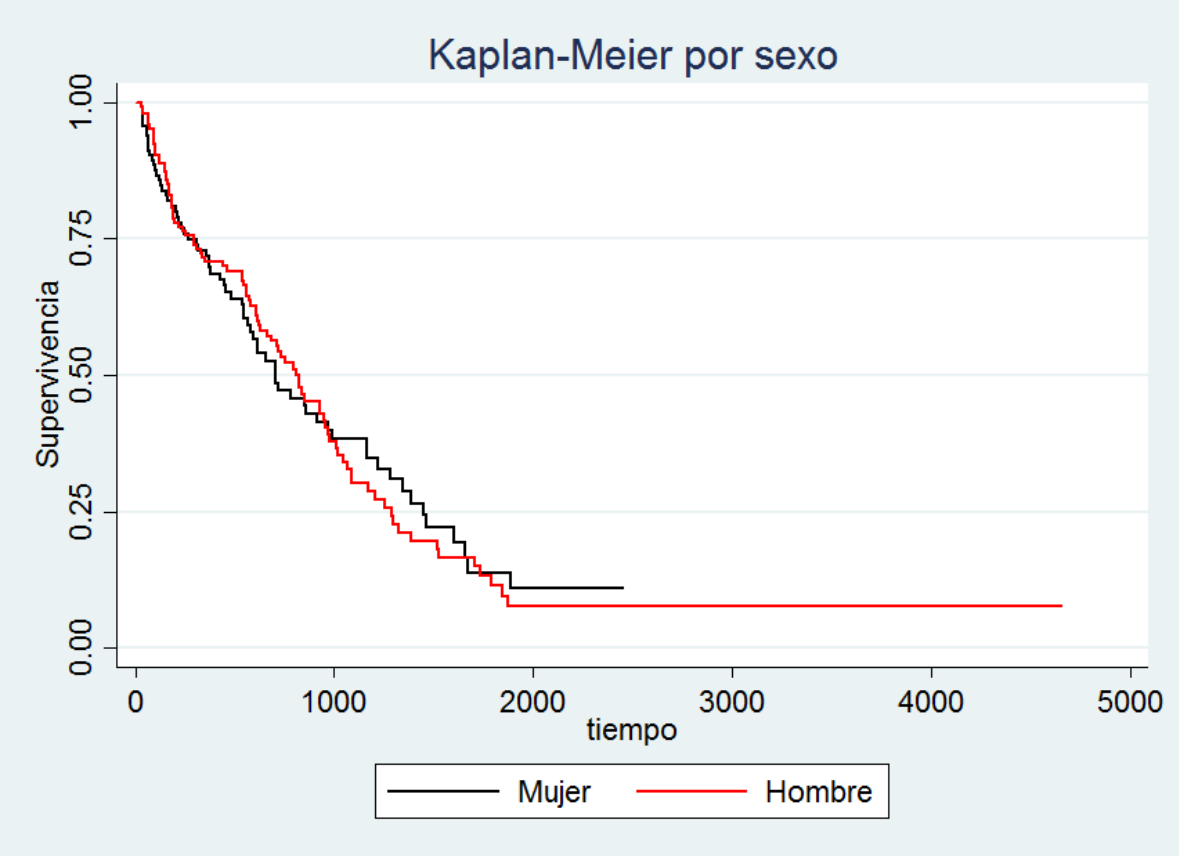
En la tabla siguiente se presentan las pruebas para la igualdad de las funciones de supervivencia de hombres y mujeres, de las cuales podemos decir que a un nivel de significancia del 5 % no existe diferencia en la supervivencia por sexo.

Tabla 4.4: Pruebas de hipótesis de igualdad de supervivencia por sexo

	Ji-cuadrada	gl	p-value
Log rank	0.02	1	0.8897
Wilcoxon	0.2	1	0.6545

En la gráfica siguiente se presentan las curvas de supervivencia para hombres y mujeres, se aprecia que al comienzo del estudio la supervivencia de las mujeres está por debajo de los hombres y alrededor de los mil días cambia y ahora se encuentran por encima.

Figura 4.2: Estimador Kaplan-Meier por sexo



4.1.2. Por consumo de calorías

A continuación se muestran los resultados, las diferentes estadísticas y curvas de supervivencia por consumo de calorías, así como sus comparaciones.

Tabla 4.5: Media del tiempo de supervivencia por consumo de calorías

Consumo de calorías	Sujetos	Eventos	Mediana	Error estándar	Intervalo 95 %	
Normal	243	134	626	52.294	560	794
Alto	121	26	1286	126.4902	824	1708

En caso contrario al sexo, se puede observar un diferencia significativa en la mediana. Las personas con un consumo normal de calorías en promedio necesitan 626 días para presentar la falla y las personas con un consumo alto, requieren 1286 días, es decir, más del doble de tiempo que una persona con consumo normal.

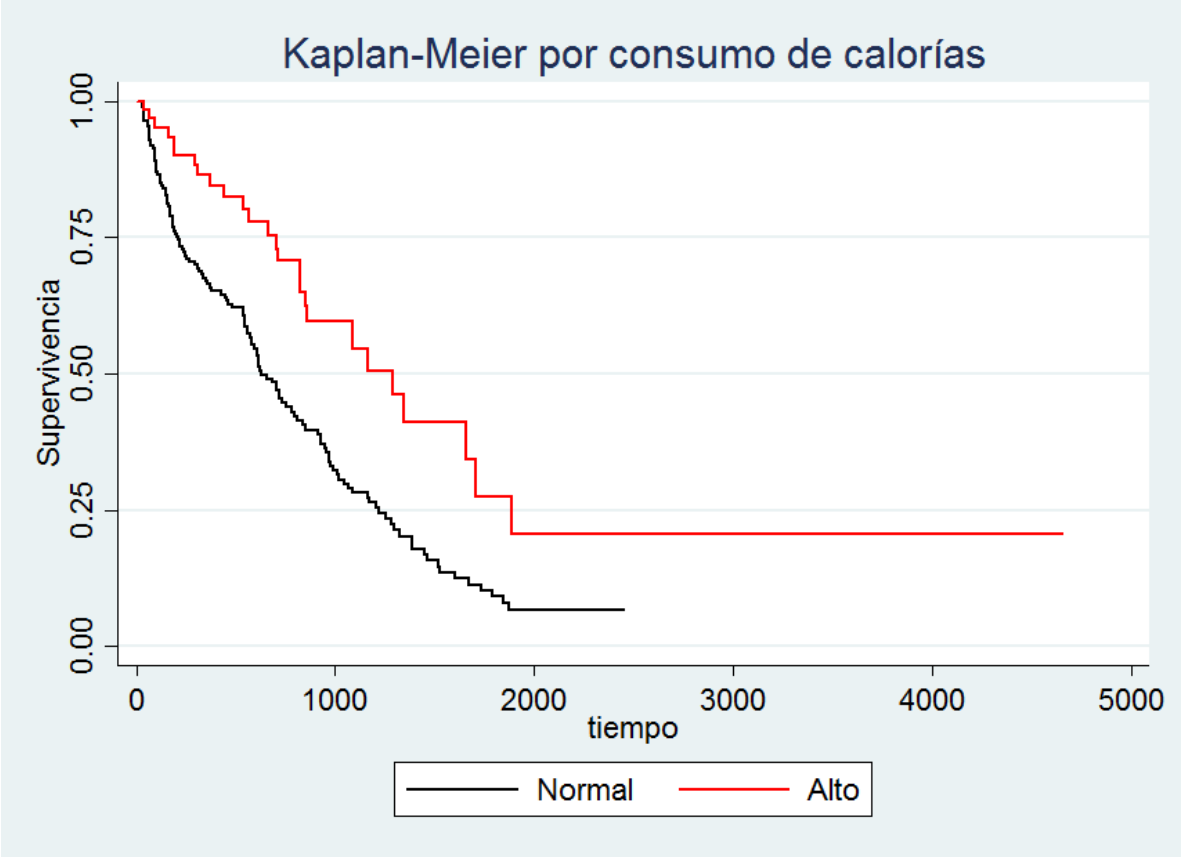
Igualmente es necesario realizar las pruebas de hipótesis correspondiente, para afirmar si existe diferencia entre la supervivencia dependiendo del consumo de calorías.

En la tabla siguiente se muestra que a un nivel de signifancia del 5 %, la función de supervivencia es distintas para personas con un consumo normal que para una con un consumo alto de calorías.

Tabla 4.6: Pruebas de hipótesis de igualdad de supervivencia por consumo de calorías

	Ji-cuadrada	gl	p-value
Log rank	12.47	1	0.0004
Wilcoxon	11.73	1	0.0006

Figura 4.3: Estimador Kaplan-Meier por consumo de calorías



La gráfica anterior muestra diferencias significativas entre las funciones de supervivencia por consumo de calorías, lo cual reafirma lo dicho por las pruebas de hipótesis.

4.1.3. Por peso al nacer

A continuación se muestran los resultados obtenidos, las curvas de supervivencia y las distintas estadísticas por peso al nacer.

Tabla 4.7: Media del tiempo de supervivencia por peso al nacer

Peso al nacer	Sujetos	Eventos	Mediana	Error estándar	Intervalo 95 %	
Bajo	22	12	540	67.738	146	
Normal	197	121	735	73.021	614	930
Alto	50	27	924	126.57	604	1283

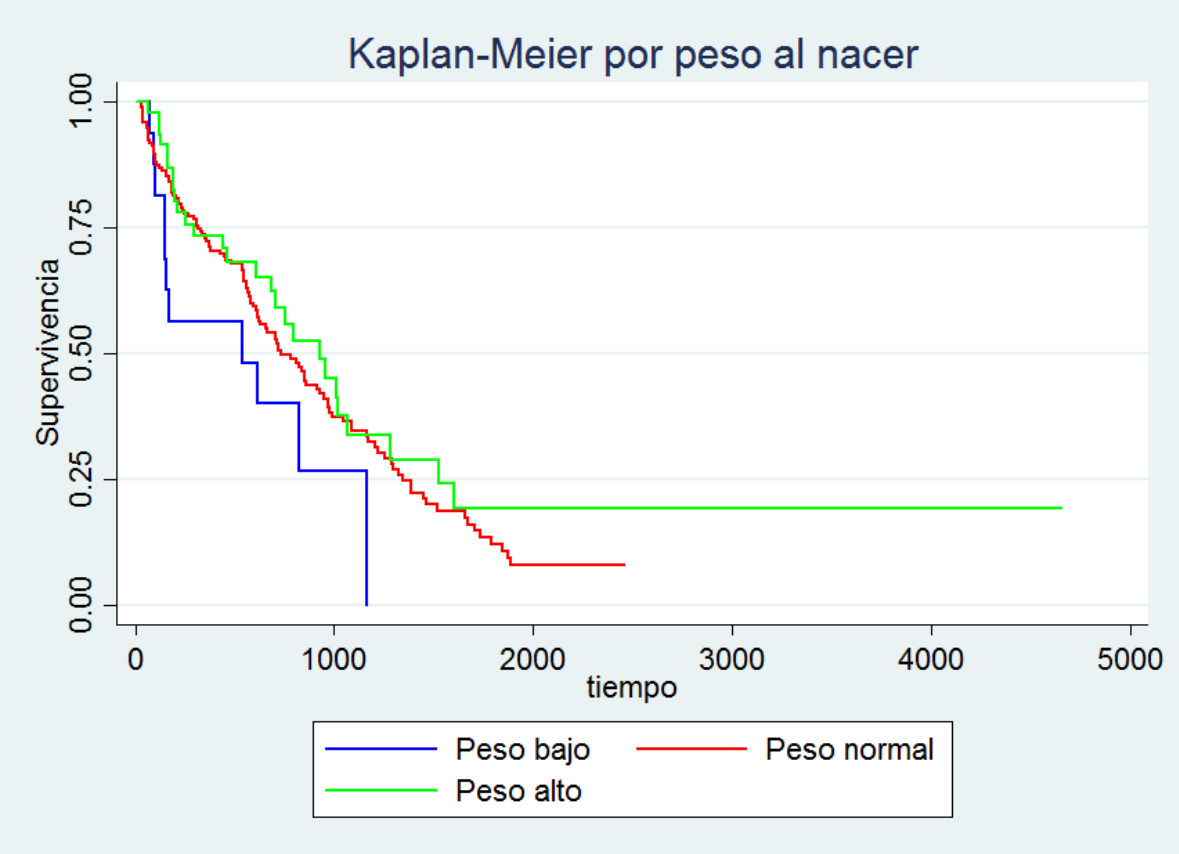
En la tabla siguiente se puede apreciar que con un nivel de significancia del 5 %, no existe evidencia para decir que la supervivencia es distinta por el peso al nacer.

Tabla 4.8: Pruebas de hipótesis de igualdad de supervivencia por peso al nacer

	Ji-cuadrada	gl	p-value
Log rank	3.63	2	0.1632
Wilcoxon	3.74	2	0.1541

La gráfica siguiente reafirma lo que muestran las pruebas de hipótesis, puesto que no se observa una diferencia significativa entre las funciones de supervivencia, lo que nos haría pensar que el peso al nacer no influye en si el paciente presenta la falla o no.

Figura 4.4: Estimador Kaplan-Meier por peso al nacer



4.1.4. Por actividad física

De igual forma que para las variables anteriores, se presentan a continuación la media de supervivencia por actividad física, así como las curvas de supervivencia y pruebas de hipótesis.

En la tabla siguiente se muestra el tiempo medio que requiere un individuo para disminuir el 5 % de su índice de masa corporal estandarizado inicial dependiendo de si realiza actividad física o no. En la cual se puede observar que no existe una diferencia siginificativa en el tiempo medio dependiendo si el sujeto realiza actividad física, puesto que el tiempo medio para un individuo que realiza actividad física es de 794 días en comparación con los 626 días que requiere una persona que no realiza.

Tabla 4.9: Media del tiempo de supervivencia por actividad física

Realiza actividad física	Sujetos	Eventos	Mediana	Error estándar	Intervalo 95 %	
No	33	19	626	42.8815	310	861
Sí	230	138	794	80.8938	665	966

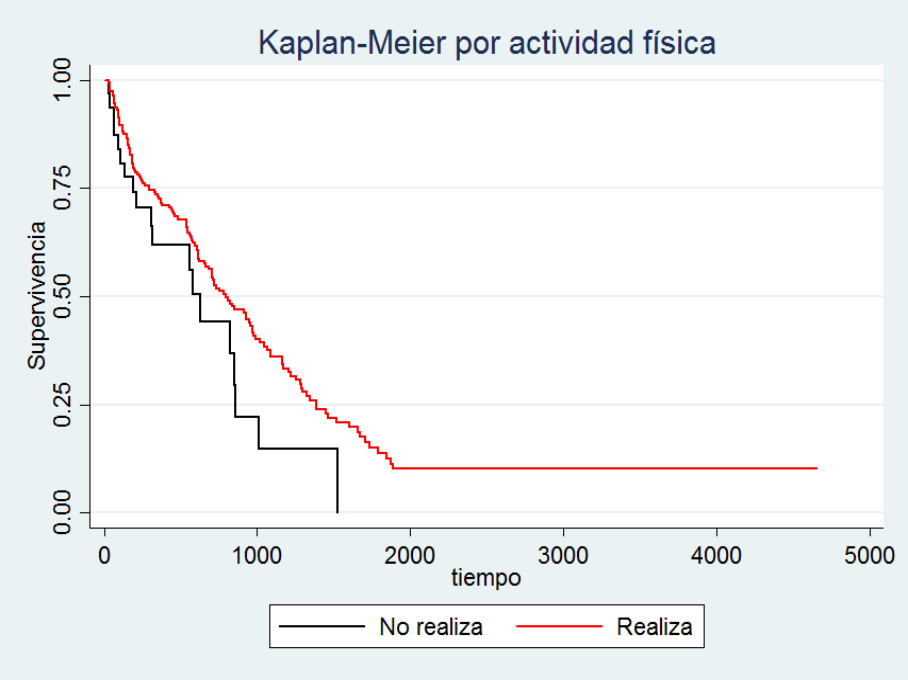
En el cuadro siguiente se muestran los resultados de las pruebas de hipótesis de igualdad de supervivencia por actividad física, de las cuales podemos afirmar que a un nivel de significancia del 5 % no existe diferencia en la supervivencia.

Tabla 4.10: Pruebas de hipótesis de igualdad de supervivencia por actividad física

	Ji-cuadrada	gl	p-value
Log rank	3.03	1	0.0816
Wilcoxon	1.96	1	0.162

La gráfica siguiente reafirma lo que muestran las pruebas de hipótesis, puesto que no se observa una diferencia significativa entre las funciones de supervivencia, lo que haría pensar que la actividad fisica no influye en que el joven disminuya el cinco por ciento de su índice de masa corporal estandarizado o no.

Figura 4.5: Estimador Kaplan-Meier actividad física



4.1.5. Por consumo de calorías y sexo

Por último analizamos una combinación que resulta interesante, sexo y consumo de calorías, puesto que es necesario saber a qué género afecta más una dieta con contenido calórico por encima de lo recomendado.

Tabla 4.11: Media del tiempo de supervivencia por consumo de calorías y sexo

Sexo-Consumo	Sujetos	Eventos	Mediana	Error estándar	Intervalo 95 %	
Mujer-Normal	104	60	541	75.9472	315	656
Mujer-Alto	58	8	1660	19.647	1165	.
Hombre-Normal	139	74	794	97.3833	609	966
Hombre-Alto	63	18	821	28.1319	438	1286

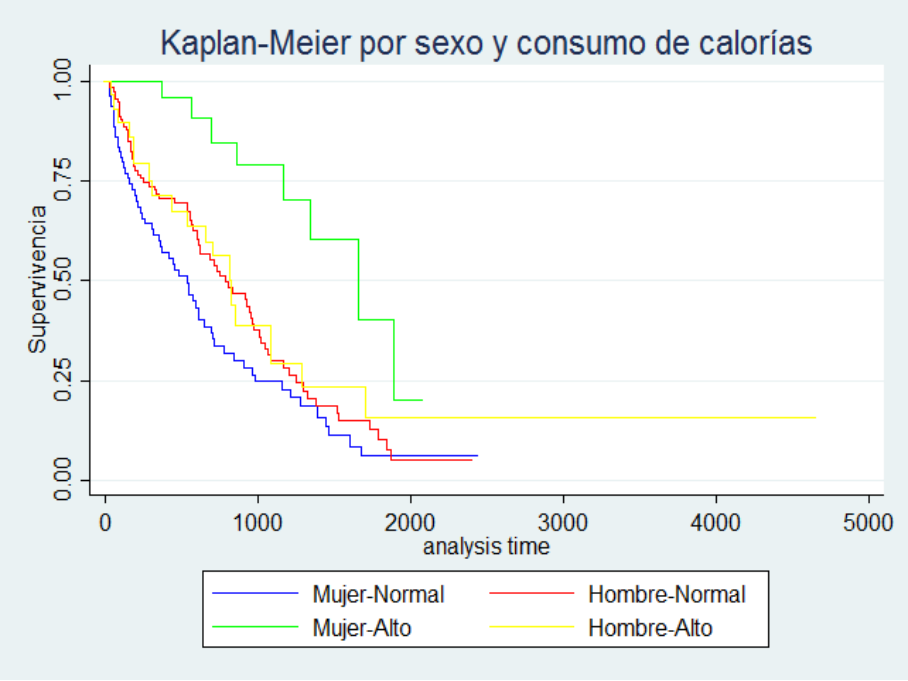
En la tabla anterior se puede apreciar que las que requieren menor tiempo para disminuir el 5 % de su índice de masa corporal estandarizado inicial son las mujeres con un consumo adecuado de calorías, pero en cambio si tienen una dieta con alto contenido calórico el tiempo que requieren es más del doble. Por otro lado se nota que en los hombre el consumo de calorías no es muy significativo, puesto que se requiere casi el mismo tiempo si lleva a cabo una dieta con alto contenido calórico o no.

En el cuadro siguiente se muestran los resultados de las pruebas de igualdad de supervivencia, en el cual podemos decir que a un nivel de significancia del 5 % existe diferencia entre la supervivencia dependiendo del sexo y el consumo de calorías.

Tabla 4.12: Pruebas de hipótesis de igualdad de supervivencia por sexo y consumo de calorías

	Ji-cuadrada	gl	p-value
Log rank	20.16	3	0.0002
Wilcoxon	23.71	3	0.0000

Figura 4.6: Estimador Kaplan-Meier por sexo y consumo de calorías



El gráfico anterior reafirma lo expuesto por las pruebas de hipótesis, ya que se puede observar una diferencia significativa entre las funciones de supervivencia, particularmente entre la supervivencia de las mujeres dependiendo del consumo de calorías.

4.2. Modelo extendido de riesgos proporcionales de Cox

El modelo de Cox que se presentará tiene como finalidad estimar el efecto de las variables de estudio sobre el tiempo de supervivencia de los pacientes. En la sección anterior se realizó un análisis exploratorio de la supervivencia de los pacientes por diversos factores, tales como sexo, actividad física y consumo de calorías, en esta sección se pretende presentar un modelo que permita explicar la supervivencia de los pacientes a partir de diversas variables, es decir, de los hábitos alimenticios de cada individuo, así como algunos factores basales.

En primera instancia se explicará la metodología para la construcción del modelo, para ello se observó la significancia una a una de las variables que se consideraron importantes. Dichas variables corresponden al grupo de macronutrientes (lípidos, carbohidratos y proteínas), así como algunas características de los pacientes (sexo, edad) y el consumo de calorías.

A continuación se presentan los distintos cuadros de modelación por cada variable continua (edad, carbohiratos, proteínas y lípidos)

Tabla 4.13: Modelo de Cox por edad

Variable	Coeficiente	Exp(coef)	Error estándar	P-value	Intervalo 95 %	
edad	-0.22476	0.7987	0.04602	0.000	-0.31495	-0.13456

Tabla 4.14: Modelo de Cox por carbohidratos

Variable	Coeficiente	Exp(coef)	Error estándar	P-value	Intervalo 95 %	
carbo_tot_dia	-0.002062	0.99793	0.000703	0.003	-0.00344	-0.000683

Tabla 4.15: Modelo de Cox por lípidos

Variable	Coeficiente	Exp(coef)	Error estándar	P-value	Intervalo 95 %	
lipid_tot_dia	-0.00789	0.99214	0.002748	0.004	-0.13273	-0.002499

Tabla 4.16: Modelo de Cox por proteínas

Variable	Coeficiente	Exp(coef)	Error estándar	P-value	Intervalo 95 %	
prote_tot_dia	-0.005626	0.99439	0.002754	0.041	-0.110233	-0.000229

De los resultados anteriores podemos decir que a un nivel de significancia del 5 % las cuatro variables son significativas, pero no sabemos si lo son conjuntamente, así que se decidió tomar el modelo con todas las variables que se consideraron importantes. Dicho modelo se presenta en la tabla siguiente

Tabla 4.17: Modelo de Cox inicial

Variable	Coeficiente	Exp(coef)	Error estándar	P-value	Intervalo 95 %	
edad	-0.21621	0.80557	0.045685	0.000	-0.305753	-0.126671
cat_kcal	-0.49617	0.60886	0.303947	0.103	-1.091899	0.099552
carbo_tot_dia	-0.001324	0.99868	0.001314	0.314	-0.003899	-0.00125
prote_tot_dia	0.016549	1.01669	0.006849	0.016	0.003125	0.029972
lipid_tot_dia	-0.013183	0.9869	0.0069	0.056	-0.026707	0.00341

Del cuadro anterior podemos decir que hay tres variables que no son significativas a un nivel del 5 %, por lo cual se tomó la decisión de sacar la variable con el p-value mayor, que es ingesta de carbohidratos y observar el comportamiento del modelo.

Tabla 4.18: Segundo modelo de Cox

Variable	Coeficiente	Exp(coef)	Error estándar	P-value	Intervalo 95 %	
edad	-0.216568	0.80528	0.045897	0.000	-0.306525	-0.126671
cat_kcal	-0.62005	0.53792	0.28089	0.027	-1.17058	-0.069516
prote_tot_dia	0.01432	1.01442	0.00656	0.027	0.001451	0.027184
lipid_tot_dia	-0.014431	0.98567	0.00683	0.035	-0.02782	0.00104

Como se puede observar del cuadro anterior, las cuatro variables son significativas, así que ya casi tenemos nuestro modelo aunque resulta interesante observar la interacción entre el consumo de calorías y el sexo, ya que en el análisis exploratorio resultó importante puesto que mostraba cierta diferencia en la supervivencia.

En el cuadro anterior se muestra que las cuatro variables anteriores son aún significativas al 5 %, además la interacción sexo y consumo de calorías de igual forma es

Tabla 4.19: Tercer modelo de Cox

Variable	Coeficiente	Exp(coef)	Error estándar	P-value	Intervalo 95 %	
edad	-0.21016	0.81045	0.04535	0.000	-0.29905	-0.12127
cat_kcal	-1.15338	0.31019	0.39564	0.004	-1.92882	-0.37795
prote_tot_dia	0.01314	1.01323	0.0064	0.04	0.00059	0.02568
lipid_tot_dia	-0.014839	0.98527	0.00666	0.026	-0.027893	0.001785
cat_kcal:sexo	1.06955	2.91407	0.43111	0.013	0.22458	1.91452

significativa.

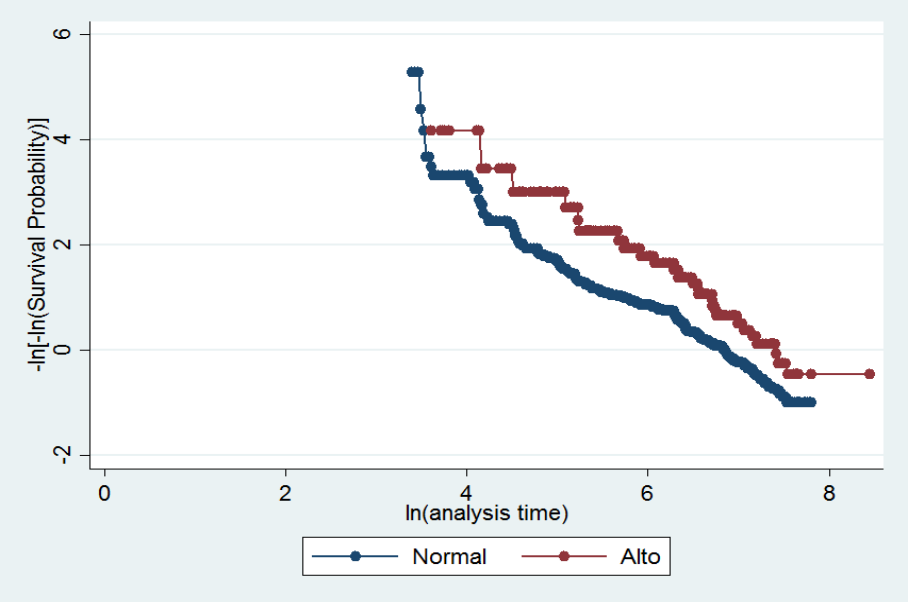
Puesto que ya tenemos nuestro modelo de riesgos proporcionales, debemos revisar si las variables cumplen los supuestos de dicha modelación. Para ello realizamos una prueba de hipótesis, de la cual se presentan los resultados en la tabla siguiente

Tabla 4.20: Pruebas del supuesto de riegos proporcionales para el tercer modelo de Cox

Variable	rho	Ji-cuadrada	p-value
edad	-0.11615	2.36	0.1242
cat_kcal	0.21003	6.65	0.0099
sexo:cat_kcal	-0.15495	3.73	0.0534
prote_tot_dia	-0.04894	0.36	0.5458
lipid_tot_dia	0.03175	0.18	0.668
global		9.8	0.0811

Del cuadro anterior podemos decir que a un nivel de significancia del 5 % la variable que no cumple el supuesto de riegos proporcionales es el consumo de calorías (cat_kcal), puesto que la hipótesis nula es que los riesgos entre las poblaciones que define cada variable son proporcionales. Dado que dicha variable es de suma importancia para explicar el riesgo, se tomó la decisión de observar a partir de qué punto se nota que cumple riesgos proporcionales, para esto se realizó una gráfica con la supervivencia transformada mediante la diferencia de logaritmos, dicha gráfica se presenta a continuación

Figura 4.7: Modelo de Cox por consumo de calorías transformado



Se puede observar que a partir de los 148 (ln(5)) días, la variable cumple el supuesto de proporcionalidad,puesto que las gráficas de la función de supervivencia comienzan a separarse a partir de tal punto.

Por lo cual se creó una nueva variable dependiente del tiempo cat_kcal_t, la cual comienza a intervenir a partir de el día 148. Dicha variable se creó a partir de la interacción de la variable cat_kcal con la función de cambio de punto siguiente

$$g(t) = \begin{cases} 1 & si \quad t \geq 148 \\ 0 & si \quad t < 148 \end{cases}$$

es decir, cat_kcal_t = cat_kcal$*g(t)$.

Por lo cual al ingresar la nueva variable se obtuvo un nuevo modelo que es

Se puede observar que las cuatro variables y la interacción son significativas, ahora de nueva cuenta debemos verificar que éstas cumplan el supuesto de riesgos proporcionales, para ello se presentan los resultados en la siguiente tabla

Por lo cual podemos afirmar que a un nivel de significancia del 5 % cada variable y de manera global el modelo cumple el supuesto de riesgos proporcionales.

En el siguiente gráfico se muestra la supervivencia ajustada mediante el modelo de

Tabla 4.21: Modelo de Cox final ajustado

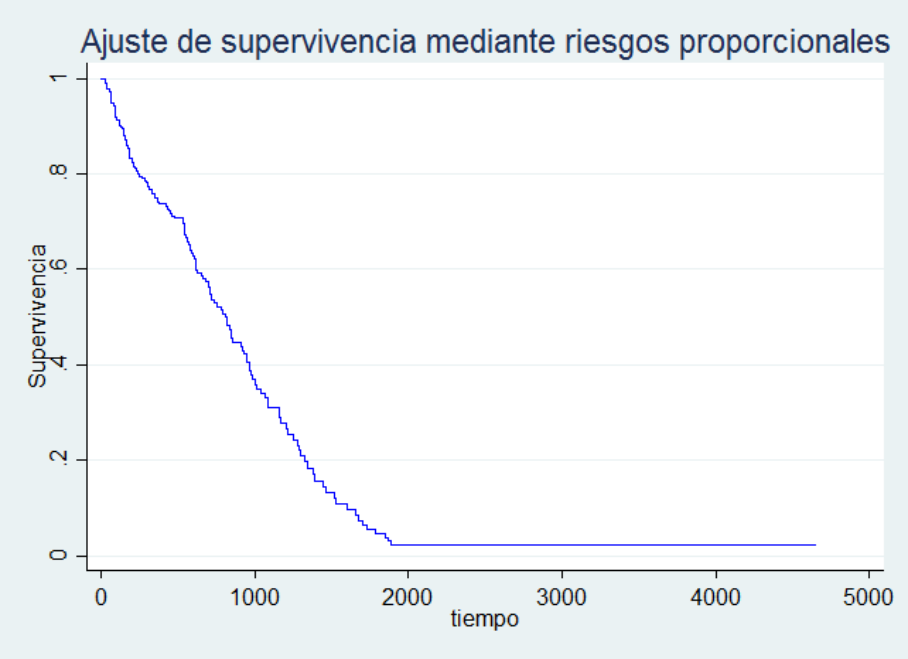
Variable	Coeficiente	Exp(coef)	Error estándar	P-value	Intervalo 95 %	
edad	-0.21299	0.80816	0.04528	0.000	-0.30174	-0.12425
cat_kcal_t	-0.81634	0.44205	0.39438	0.038	-1.58932	-0.04337
prote_tot_dia	0.01284	1.01292	0.00641	0.045	0.00027	0.02541
lipid_tot_dia	-0.01721	0.98294	0.00666	0.01	-0.03027	0.00416
cat_kcal_t:sexo	0.98375	2.6745	0.44258	0.026	0.11632	1.85119

Tabla 4.22: Pruebas del supuesto de riegos proporcionales para el modelo de Cox final ajustado

Variable	rho	Ji-cuadrada	p-value
edad	-0.11593	2.36	0.1245
cat_kcal_t	0.13713	2.87	0.092
sexo:cat_kcal_t	-0.12415	2.41	0.1206
prote_tot_dia	-0.04439	0.31	0.578
lipid_tot_dia	0.05379	0.53	0.4651
global		6.49	0.2615

riesgos proporcionales de Cox

Figura 4.8: Ajuste de supervivencia mediante riesgos proporcionales



4.3. Análisis de residuos

En esta sección se presentan los residuos del ajuste mediante el modelos de riesgos proporcionales de Cox, así como su interpretación.

4.3.1. Residuos escalados de Schoenfeld

Los residuos escalados de Schoenfeld nos muestran si cada variable cumple la condición de riesgos proporcionales. Al ajustar una línea a la gráfica de los residuos contra el tiempo, dicha línea debe verse aproximadamente horizontal.

En las siguientes gráficas se muestran los residuos escalados de Schoenfeld para cada variable

Figura 4.9: Residuos de Schoenfeld para edad

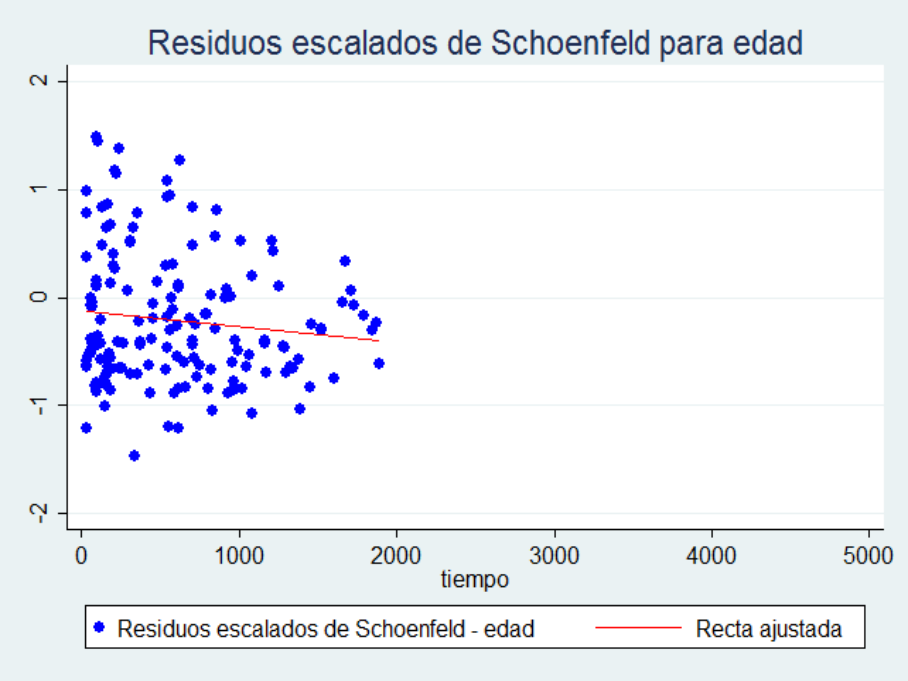


Figura 4.10: Residuos de Schoenfeld para calorías

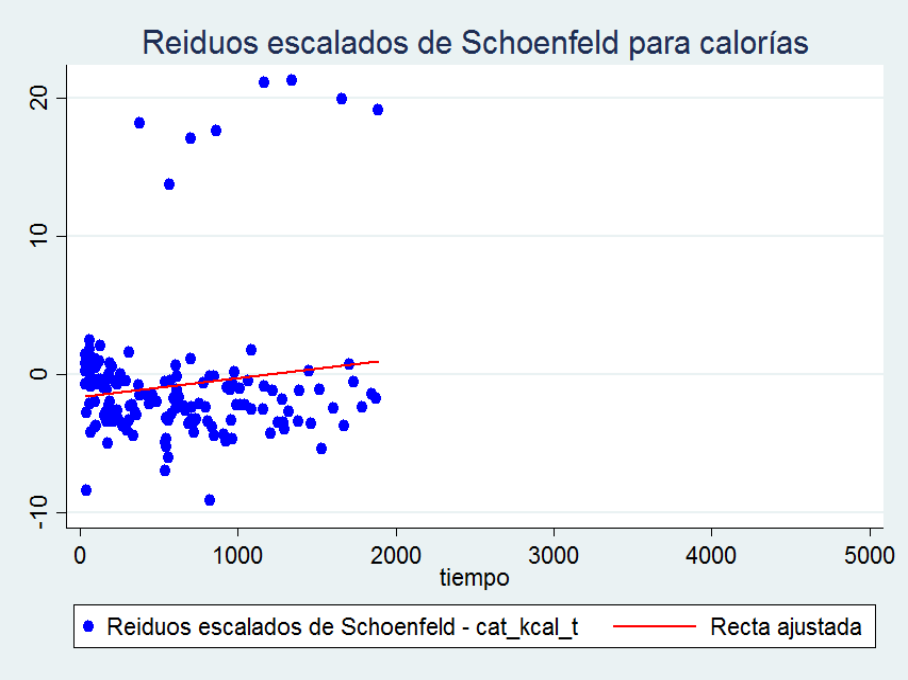


Figura 4.11: Residuos de Schoenfeld para proteínas

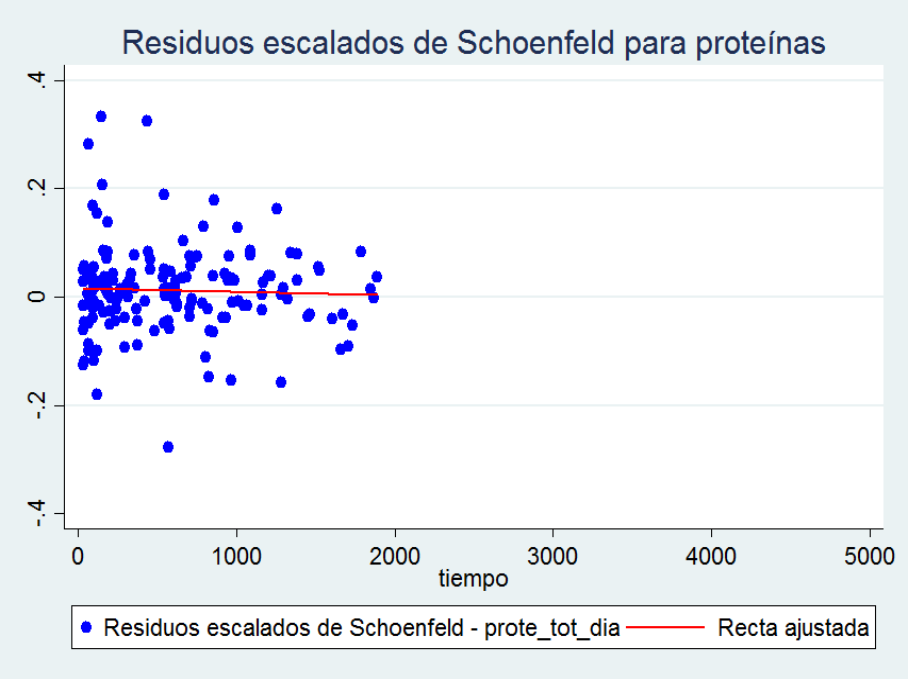
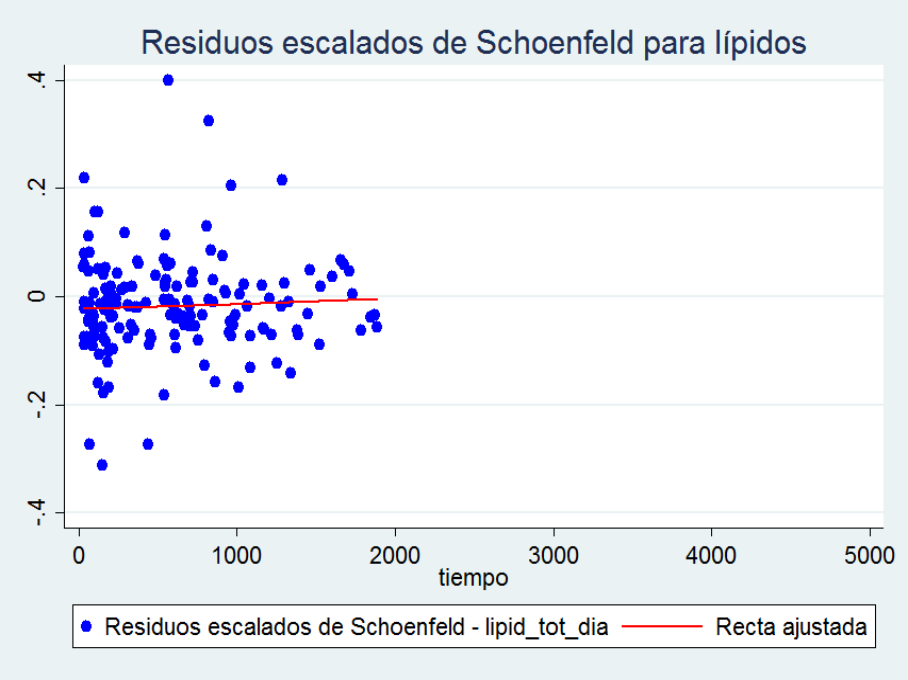


Figura 4.12: Residuos de Schoenfeld para lípidos



De las gráficas anteriores podemos decir que la pendiente de la recta es aproximadamente cero por lo cual las variables cumplen el supuesto de riesgos proporcionales, también existe una relación importante entre la apariencia de la recta con el p-value de la prueba del supuesto de riesgos proporcionales de la sección anterior, ya que entre menor sea el p-value la pendiente de la recta se ve más alejada de cero, en caso contrario cuando el p-value es mayor la recta de ajuste parece ser horizontal. Por lo cual la variable cat_kcal_t cuyo p-value (0.092) es el menor de todos, muestra una recta de ajuste que se ve con una pendiente más pronunciada con respecto a las otras variables, caso contrario con las proteínas con el p-value mayor (0.578) la recta aproximadamente parece ser horizontal.

4.3.2. Residuos de martingalas

Los residuos de martingalas muestran si la variable con la que se comparan tiene una relación lineal con el riesgo. Puesto que la comparación de los residuos con variables dicotómicas resulta poco útil, ya que para decir que existe dicha relación lineal los residuos deben verse distribuidos al azar, por lo cual sólo se analizarán los casos de variables continuas.

A continuación se muestran los residuos de martingalas comparados con cada variable continua

Figura 4.13: Residuos de martingalas por edad

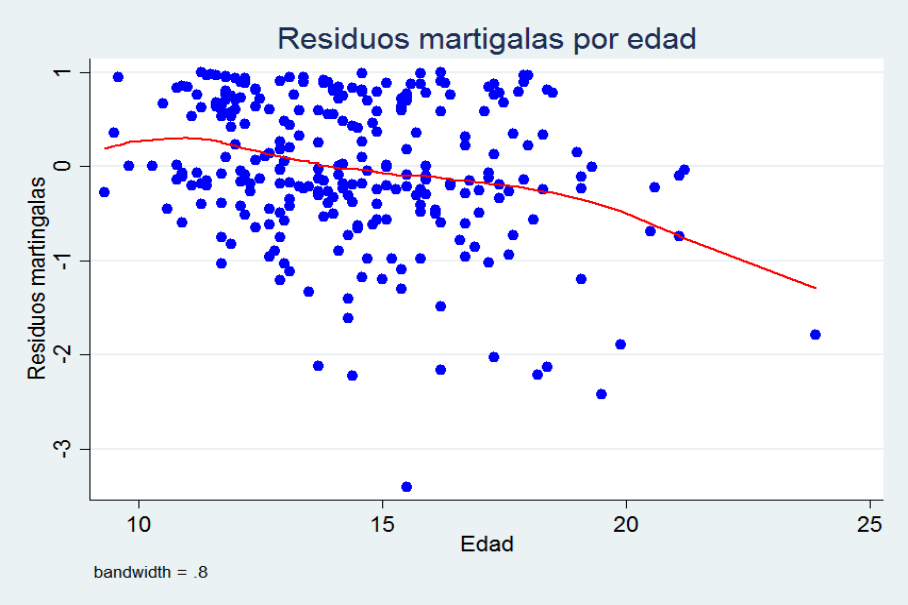


Figura 4.14: Residuos de martingalas por proteínas

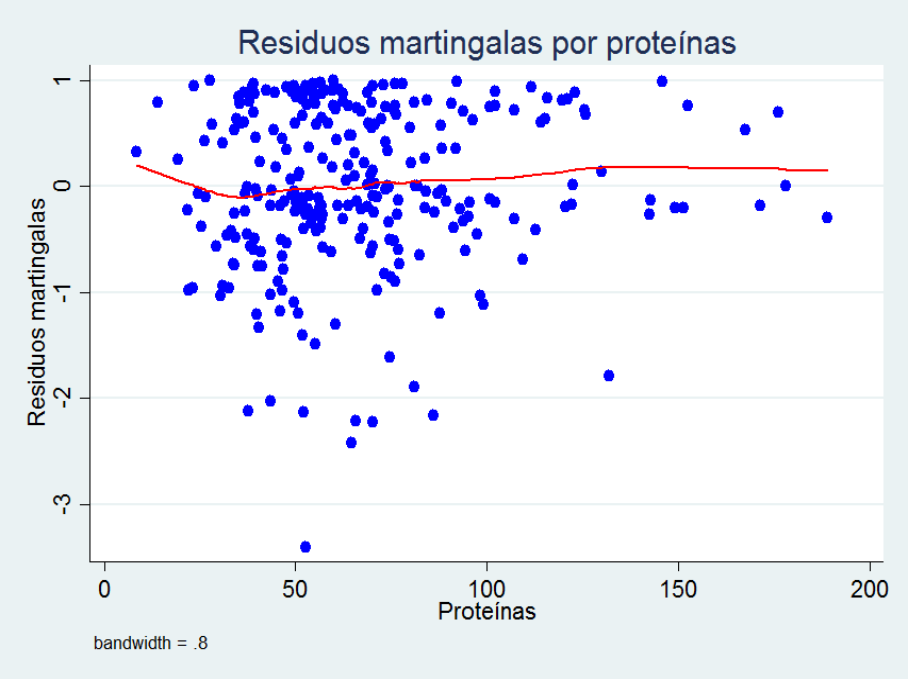
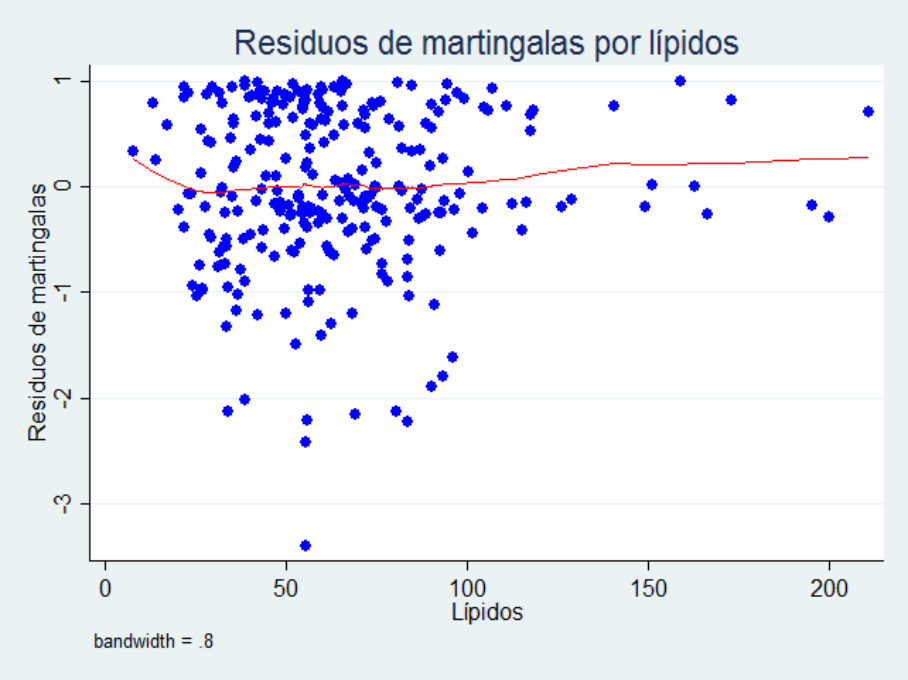


Figura 4.15: Residuos de martingalas por lípidos

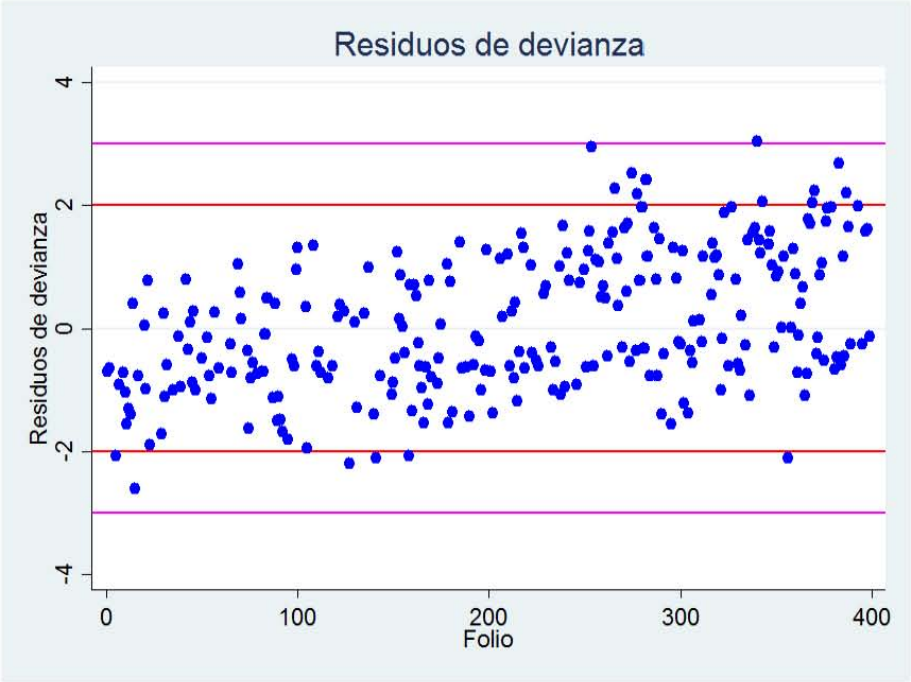


De las gráficas anteriores podemos concluir que las tres variables continuas, (edad, proteínas y lípidos) tienen una relación lineal con el riesgo, ya que se observa que los residuos de martingalas se distribuyen de manera aleatoria.

4.3.3. Residuos de devianza

Los residuos de devianza son una transformación de los residuos de martingalas y sirven para evitar el sesgo a la derecha de éstos, además de servir para detectar observaciones con residuos grandes (mayores a 2 o 3 en valor absoluto), puesto que ya se utilizó los residuos de martingalas para verificar linealidad de las variables, usaremos los residuos de devianza para verificar si tenemos observaciones con residuos grandes, lo cual se puede observar en la gráfica siguiente, ya que son menores que 3 en valor absoluto.

Figura 4.16: Residuos de devianza



4.3.4. Dfbetas

Los dfbetas sirven para reconocer qué observaciones influyen más o cuáles son datos atípicos, es decir los extremos que se presentan por variable

En seguida se presentan las dfbetas por cada variable del modelo

Figura 4.17: Influencia por edad

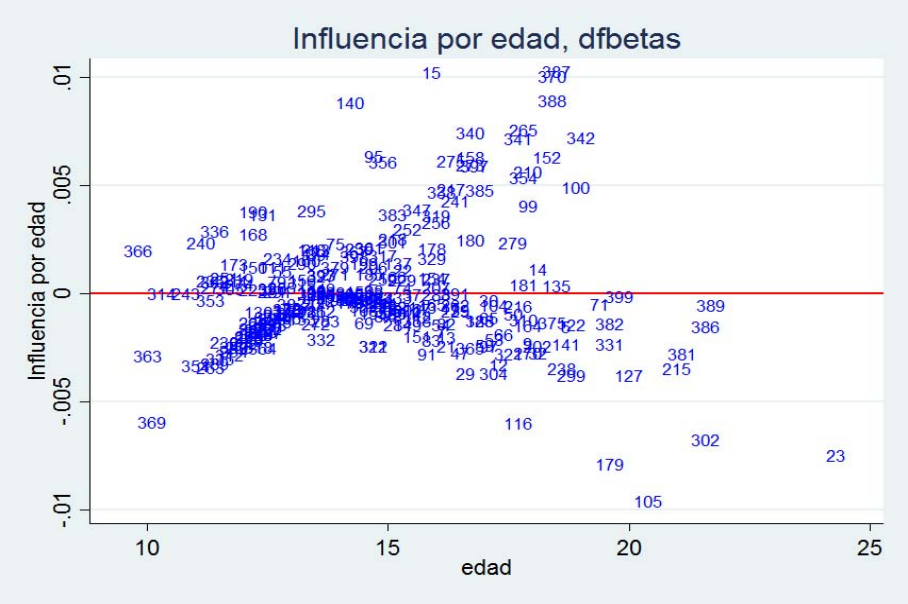


Figura 4.18: Influencia por calorías

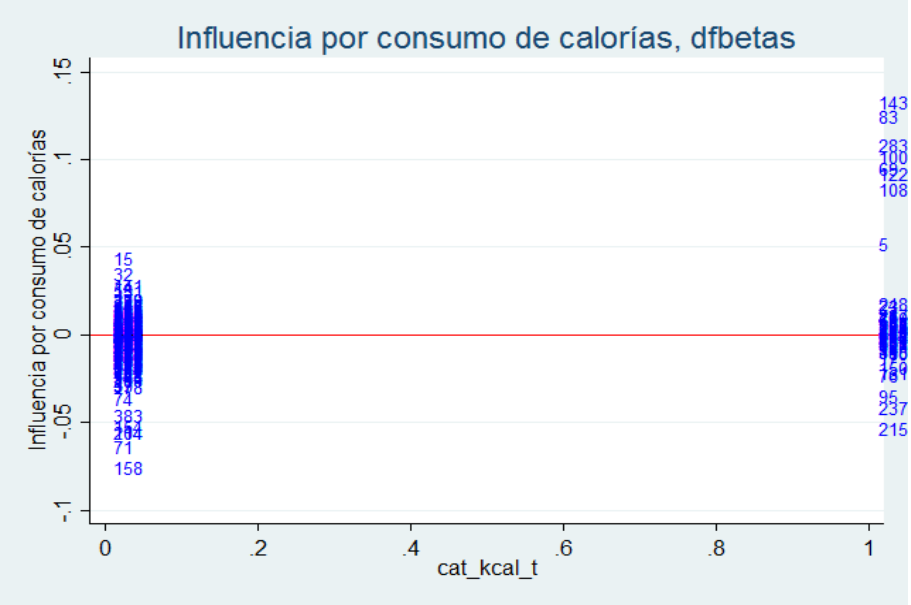


Figura 4.19: Influencia por proteínas

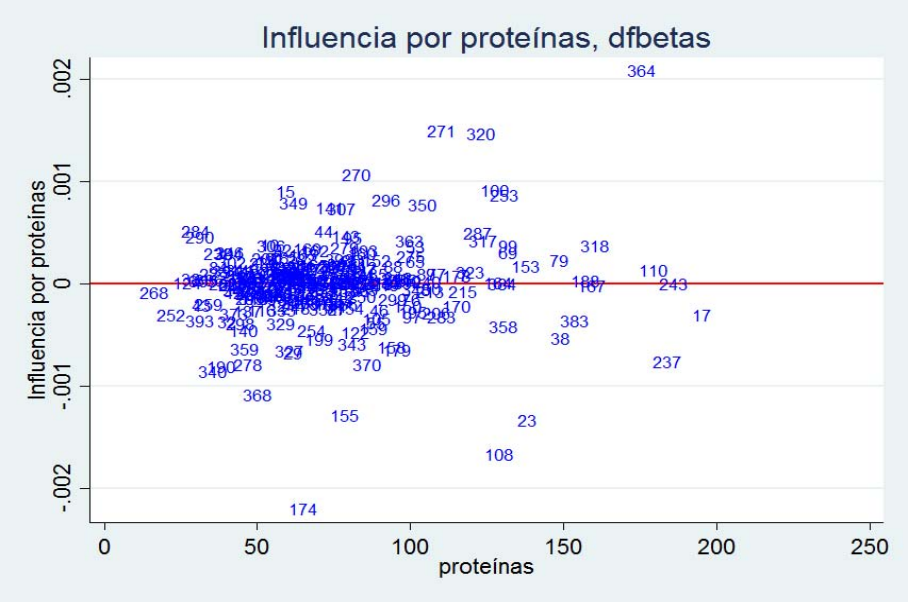
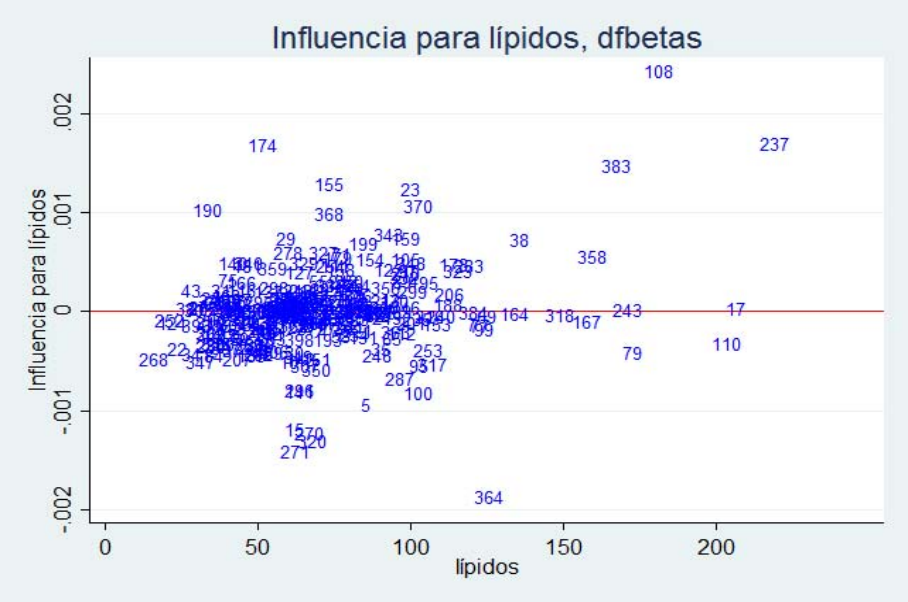


Figura 4.20: Influencia por lípidos



De las gráficas anteriores se puede observar que los valores de las dfbetas no están muy alejados de cero, aunque si tomamos un criterio estricto, en la edad los datos extremos son aquellas personas con una edad aproximada de 20 años.

4.4. Conclusiones

Estimador Kaplan-Meier

Se observó que el tiempo promedio que requiere un sujeto para disminuir el 5 % de su índice de masa corporal estandarizado (z-score) es de 784 días, aproximadamente 2 años, lo cual podría deberse a que la falta de compromiso de los jóvenes, puesto que algunos dejaban de asistir a las citas medicas y regresaban después mucho tiempo. También se observó que el género no influye en el tiempo que requiere una persona para lograr disminuir dicho porcentaje, lo que influye en gran medida es si la persona respeta el límite de calorías que puede consumir por día. Además se puede afirmar que la supervivencia de los hombres no se ve afectada si consume más calorías de lo recomendado, en cambio en las mujeres cambia radicalmente el tiempo requerido para disminuir el 5 % del índice de masa corporal estandarizado si consume más calorías de las recomendadas por día.

Modelo de riesgos proporcionales de Cox

Tabla 4.23: Coeficientes del modelo de Cox final ajustado

Variable	Coeficiente	Exp(coef)	P-value
edad	-0.21299	0.80816	0.000
cat_kcal_t	-0.81634	0.44205	0.038
prote_tot_dia	0.01284	1.01292	0.045
lipid_tot_dia	-0.01721	0.98294	0.01
cat_kcal_t:sexo	0.98375	2.6745	0.026

La edad es un factor importante puesto que una persona tiene menos probabilidad de disminuir el 5 % de su índice de masa corporal estandarizado inicial con respecto a la probabilidad de otra un año menor, esto es el riesgo de presentar la falla de una persona es 0.80816 veces el riesgo de una persona un año menor, es decir un año adicional disminuye aproximadamente en un 20 % el riesgo.

En cuanto al consumo de calorías, una persona con un consumo alto es menos propensa a disminuir el 5 % de su índice de masa corporal estandarizado inicial que una persona con un consumo normal, esto es, el riesgo de una persona con un consumo alto de calorías es 0.44205 veces el riesgo de una persona que tiene un consumo normal.

Además una mujer con un consumo alto de calorías es menos propensa a disminuir el 5 % de su índice de masa corporal estandarizado inicial que un hombre con el mismo tipo de consumo, puesto que el riesgo de sufrir la falla de un hombre con una dieta alta en calorías es 2.6724 veces el riesgo de una mujer con el mismo tipo de consumo. Este fenómeno puede deberse a los efectos mediados a través del receptor de estrógenos alfa (Erα), ya que la activación de este receptor regula la ingesta de alimentos, la homeostasis de la glucosa y el gasto energético. Las mujeres experimentan fluctuaciones en la ingesta de alimentos dependiendo de la fase de su ciclo menstrual.

De los dos macronutrientes que se analizaron podemos decir que el consumo de un gramo más de proteínas es benéfico para disminuir el 5 % del índice de masa corporal estandarizado inicial, caso contrario con los lípidos, puesto que el consumo de un gramo más disminuye el riesgo de presentar la falla. En cuanto al riesgo, un gramo adicional de proteínas lo aumenta un 1.292 % y un gramo adicional de lípidos lo disminuye 1.706 %. Esto muestra que la limitación de la ingesta de grasas disminuye el riesgo de obesidad, por ser un nutriente con mayor densidad energética, lo que significa que dicha limitación hace más fácil la disminución de energía total. La cantidad total de grasas como el patrón de ácidos grasos debe ser considerada en las recomendaciones dietéticas. Por lo que se sugiere que además de la limitación de la ingesta de grasa debe haber una modificación del patrón de ácidos grasos a favor de los ácidos grasos insaturados, particularmente de ácidos grasos n-3 de cadena larga. El incremento de la ingesta de proteínas en la pérdida de peso tiene beneficios en parte por el efecto de ser un nutriente que ocasiona mayor sensación de saciedad funcionando como un excelente regulador de la ingesta calórica. Una mayor ingesta de proteína durante la pérdida de peso previene la pérdida de masa corporal magra y por lo tanto puede mejorar la sensibilidad a la insulina.

Bibliografía

Cleves, M., Gould, W., and Marchenko, Y. V. (2010). *An introduction to survival analysis using stata*, volume 3. Stata Press.

Klein, J. P. and Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kleinbaum, D. G. and Klein, M. (2006). *Survival analysis: a self-learning text*. Springer Science & Business Media.

OMS., S. d. I. T. (2003). *Dieta, nutrición y prevención de enfermedades crónicas*. OMS (Organización Mundial de la Salud).

Otero Lamas, B. (2012). *Nutrición*. Red Tercer Milenio.

Therneau, T. M. and Grambsch, P. M. (2013). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.