00365

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

POSGRADO EN CIENCIAS MATEMATICAS

FACULTAD DE CIENCIAS

ANALISIS DE SOBREVIVENCIA APLICADO A UN ESTUDIO DE PREECLAMPSIA

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE:

MAESTRA EN CIENCIAS MATEMATICAS

P R E S E N T A :

JESICA HERNANDEZ ROJANO

DIRECTORA DE TESIS:

DRA. SILVIA RUIZ – VELASCO ACOSTA

MEXICO, D. F. **ENERO 2005**

Autorizo a la Dirección General de Bibliotecas de la UNAM a difundir en formato electrónico e impreso el contenido de mi trabajo recepcional.

NOMBRE: Jesica Hernández Rojano

FECHA: 25 - Enero - 2005

FIRMA: Jesica Hernández Rojano

INDICE

INTRODUCCIÓN

A mediados del 2002 el Instituto Nacional de Perinatología acudió al IIMAS a exponer un proyecto de investigación titulado "Ensayo Clínico de suplementación con L-arginina y vitaminas antioxidantes para prevención de desarrollo de preeclampsia", cuyo objetivo era evaluar la eficacia de la suplementación con L-arginina y vitaminas antioxidantes, administradas en una barra alimenticia, en la incidencia de preeclampsia en una población de alto riesgo.

Al interesarnos el trabajo de investigación del INPer decidimos participar en la parte estadística, necesaria para interpretar los datos con los cuales se contaba. De ahí es de donde surgió el presente trabajo. Uno de los objetivos del estudio era el evaluar si el tiempo de gestación era diferente si se tomaba el suplemento. Nuestro objetivo en ese momento era aplicar técnicas de análisis de sobrevivencia en la base de datos completa, para poder obtener conclusiones que ayudaran al INPer en su investigación.

Cabe mencionar que al momento de acudir al IIMAS, el INPer se encontraba en pleno desarrollo de la investigación y contaba sólo con una pequeña parte de la base de datos, que nos proporcionó en ese momento, prometiendo actualizarla conforme la fueran completando. Sin embargo, esto nunca sucedió, por lo que tuvimos que trabajar con los datos proporcionados y no se pudo cumplir el objetivo principal, quedando este trabajo únicamente como una aplicación, sin poder obtener conclusiones válidas para la población para la cual estaba dirigido el estudio.

En el capítulo 1 se presenta un resumen de las técnicas de Análisis de Sobrevivencia utilizadas en el presente trabajo. En el capítulo 2, Aplicaciones, se presentan los modelos obtenidos y sus interpretaciones, además de ciertas técnicas para saber si tales modelos son adecuados. En la parte de Conclusiones, se presentan, además, las limitaciones que se tuvieron al desarrollar el presente trabajo y en qué manera lo afectaron.

1 ANÁLISIS DE SUPERVIVENCIA

Se le llama Análisis de Supervivencia al análisis de datos que miden el tiempo que pasa desde el *tiempo cero* u *origen* hasta la ocurrencia de cierto evento o *punto de término.* En investigaciones médicas, el origen generalmente corresponde al momento en que el individuo es reclutado en un estudio experimental, como lo es un ensayo clínico para comparar dos o más tratamientos. A su vez, puede coincidir con el diagnóstico de una condición en particular, el comienzo de un régimen de tratamiento o la ocurrencia de algún evento. Si el punto de término es la muerte del paciente, los datos resultantes son literalmente tiempos de supervivencia; sin embargo, pueden obtenerse datos similares cuando el punto de término no es fatal, tal como el alivio de la enfermedad, la recurrencia de síntomas o la falla de algún sistema o mecanismo (comúnmente, para evaluaciones teóricas, al evento considerado se le llama *muerte* o *falla* y al momento en el que ocurre esa muere o falla se le llama *tiempo de muerte o falla*).

Los métodos para analizar datos de supervivencia no están restringidos a tiempos de supervivencia de forma literal, sino que se aplican igualmente a datos que se refieren al tiempo que transcurre hasta otros puntos de término. La metodología puede ser utilizada también en datos de otras áreas de aplicación como, por ejemplo, las industriales, económicas, demográficas, etc.

1.1 CARACTERÍSTICAS ESPECIALES DE LOS DATOS DE SUPERVIVENCIA

Generalmente, los datos de supervivencia no se distribuyen simétricamente. Típicamente, un histograma de tiempos de supervivencia de un grupo de individuos similares tenderá a estar *sesgado positivamente*, es decir, tendrá una "cola" más larga a la derecha del intervalo que contiene el mayor número de observaciones. Como consecuencia, no es razonable suponer que los datos de

este tipo tengan distribución normal, como se hace en los procedimientos estadísticos estándar. Este problema puede resolverse transformando primero los datos para dar una distribución más simétrica, por ejemplo calculando logaritmos; sin embargo, un método más satisfactorio es adoptar un modelo distribucional alternativo para los datos originales.

Una segunda característica de los datos de supervivencia es que frecuentemente están *censurados*. Se dice que el tiempo de supervivencia de un individuo está censurado cuando el punto de término de interés no se ha observado para ese individuo. Esto puede suceder debido a que los datos de un estudio se analizarán en un momento en el cual el individuo no ha presentado el evento de término. Alternativamente, el status de supervivencia puede no conocerse en el momento del análisis debido a que se *perdió el seguimiento* del individuo. Como un ejemplo, supóngase que después de ser reclutado para un ensayo clínico, un paciente se muda a otra parte del país o a otro país y ya no puede ser rastreado. La única información disponible de su supervivencia es la última fecha en la cual se sabía estaba vivo, que bien puede ser la última vez que el paciente se reportó a la clínica para su revisión regular.

Un tiempo de supervivencia también puede considerarse censurado cuando se sabe que la muerte se debió a alguna causa no relacionada con el tratamiento. En muchos casos, es difícil estar seguro de que la muerte no estuvo relacionada a un tratamiento particular que el paciente ha seguido. Por ejemplo, considérese un paciente que está dentro de un ensayo clínico que compara terapias alternativas para cáncer de próstata y que muere por un accidente de tráfico. El accidente pudo ser causado por un ataque de vértigo que pudo ser un efecto secundario del tratamiento al cual el paciente estaba asignado. Si es así, la muerte está, de alguna manera, relacionada con el tratamiento. En circunstancias como esa, el tiempo de supervivencia hasta la muerte por todas las causas o el tiempo de muerte por causas diferentes a la condición primaria por la cual el paciente está siendo tratado, pueden ser sujetos a un análisis de supervivencia (aunque en el análisis original esta información se considera censurada).

En cada una de estas situaciones, un paciente que entra a un estudio en el tiempo t_0 muere en el tiempo $t_0 + t$. Sin embargo, t es desconocida, ya sea porque el individuo todavía está vivo o porque se ha perdido seguimiento de él. Si se supo por última vez que el individuo estaba vivo en el tiempo $t_0 + c$, el tiempo c es llamado tiempo de supervivencia censurado. Esta censura ocurre después de que el individuo ha entrado al estudio, es decir, a la derecha del último tiempo de supervivencia conocido y se le llama *censura por la derecha.* El tiempo de supervivencia censurado por la derecha es entonces menor al verdadero, pero desconocido, tiempo de supervivencia.

Otra forma de censura es la *censura por la izquierda,* que se da cuando el tiempo de supervivencia real de un individuo es menor que el observado. Para ilustrar este tipo de censura, considérese un estudio en el cual el interés se centra en el tiempo de recurrencia de un cáncer en particular después de que se removió quirúrgicamente el tumor primario. Tres meses después de la operación los pacientes son examinados para determinar si existe recurrencia. En este momento, se encuentra que algunos de los pacientes han tenido recurrencia. Para tales pacientes, el tiempo real de recurrencia es menor a tres meses y los tiempos de recurrencia de esos pacientes están censurados por la izquierda. La censura por la izquierda es menos común que la censura por la derecha.

Otro tipo de censura es la *censura por intervalo.* Aquí, no se sabe el tiempo exacto en que los individuos han experimentado la falla, pero se sabe que fue dentro de cierto intervalo de tiempo. Considérese el ejemplo de la recurrencia del tumor utilizado en el párrafo anterior. Si se observa que el paciente está libre de la enfermedad a los tres meses, pero si en la revisión realizada a los seis meses de la operación se encuentra la enfermedad, se sabe que el tiempo de recurrencia real de tal paciente está entre tres y seis meses, entonces se dice que el tiempo de recurrencia está *censurado en el intervalo.* Aunque podría confundirse este tipo de censura con la censura por la izquierda, nótese que esta última se refiere a una recurrencia del tumor sólo entre el "inicio" del estudio y la primera revisión, siendo censura por intervalo si esta recurrencia se presenta entre revisiones.

1.1.1 Tiempo del paciente y tiempo del estudio

En un estudio típico, no todos los pacientes son reclutados exactamente al mismo tiempo, sino que se van acumulando sobre un periodo de meses o de años. Después del reclutamiento, se hace un seguimiento de los pacientes hasta que mueren o hasta un punto del calendario que marca el fin del estudio, cuando se analizan los datos. Aunque los tiempos de supervivencia reales se observarán para cierto número de pacientes, después del reclutamiento se habrá perdido el seguimiento de algunos de ellos mientras que otros todavía estarán vivos al final del estudio. Al periodo de tiempo-calendario durante el cual un individuo está en el estudio se le conoce como *tiempo del estudio.* En la Figura 1.1 se ilustra con un diagrama el tiempo de estudio para ocho individuos en un ensayo clínico, en el cual el tiempo de entrada al estudio se representa por un "•". Los individuos 1, 4, 5 y 8 mueren (M) durante el curso del estudio, se les pierde el seguimiento a los individuos 2 y 7 (P) y los individuos 3 y 6 aún están vivos (V) al final del periodo de observación.

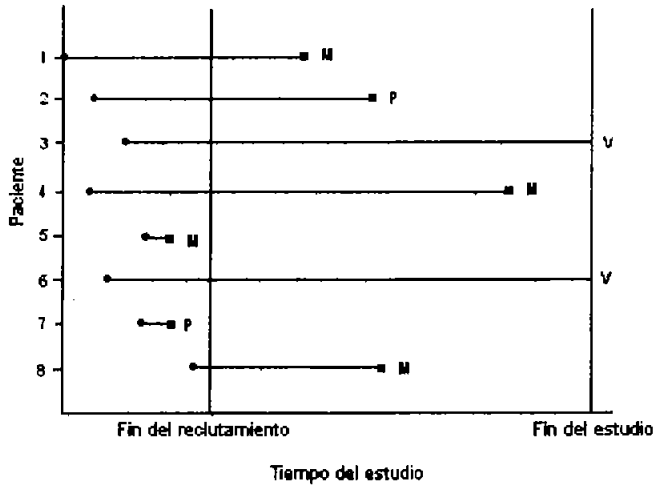

Figura 1.1. Tiempo de estudio para ocho pacientes en un estudio de supervivencia.

En lo que respecta a cada paciente, el ensayo empieza en algún tiempo t_0. El periodo de tiempo que el paciente invierte en el estudio, medido desde el tiempo origen de tal paciente, frecuentemente se conoce como *tiempo del paciente.*

El periodo de tiempo desde el origen hasta la muerte del paciente (M) es entonces el tiempo de supervivencia y ha sido registrado para los individuos 1, 4, 5 y 8. Los tiempos de supervivencia para el resto de los individuos están *censurados por la derecha.*

Una suposición importante que se hará en el análisis de datos de supervivencia censurados es que el tiempo de supervivencia real de un individuo, t, es independiente de cualquier mecanismo que cause que el tiempo de supervivencia del individuo esté censurado al tiempo c, donde $c < t$. Esto significa que si consideramos un grupo de individuos que tienen los mismos valores para las variables de pronóstico relevantes, el individuo cuyo tiempo de supervivencia está censurado al tiempo c debe ser representativo de todos los demás individuos en el grupo que han sobrevivido hasta ese tiempo. Un paciente cuyo tiempo de supervivencia está censurado será representativo de aquellos en riesgo en el tiempo de censura si el proceso de censura opera aleatoriamente. Similarmente, cuando los datos de supervivencia sean analizados en algún punto predeterminado del calendario o en un intervalo de tiempo fijo después del origen para cada paciente, el pronóstico para los individuos que aún están vivos puede ser considerado independiente de la censura, siempre y cuando el tiempo de análisis se especifique antes de que se examinen los datos. Sin embargo, esta suposición no puede hacerse si, por ejemplo, el tiempo de supervivencia de un individuo es censurado debido a que fue retirado del tratamiento como resultado de un deterioramiento en su condición física. Este tipo de censura es conocida como *censura informativa.* Debe tenerse mucho cuidado en asegurar que cualquier censura en los tiempos de supervivencia sea no informativa, porque de otra forma los métodos presentados para el análisis de datos de supervivencia no serán válidos.

1.2 Funciones de supervivencia y funciones de riesgo.

Definición 1.1. El tiempo real de supervivencia de un individuo, t, puede ser considerado como el valor de una variable aleatoria T, que puede tomar cualquier

valor en $[0,\infty)$. A T se le llama *variable aleatoria de supervivencia.*

Supóngase que T tiene densidad de probabilidad f y función de distribución acumulada F. Entonces

$$F(t) = P(T < t) = \int_0^t f(u)\,du \tag{1.1}$$

Definición 1.2. La *función de supervivencia*, S, se define como $S(t) = 1 - F(t)$.

Esto significa que

$$S(t) = P(T \geq t) = \int_t^\infty f(u)\,du \tag{1.2}$$

La función de supervivencia S es monótonamente decreciente sobre su soporte $[0,\infty)$. Además, S satisface $S(0) = 1, S(\infty) = 0$.

De hecho, cualquier función monótona decreciente S con soporte $[0,\infty)$ y con $S(0) = 1, S(\infty) = 0$ es una función de supervivencia de alguna variable aleatoria de supervivencia. Tal variable aleatoria es la que tiene función de densidad de probabilidad $f(u) = -\dfrac{dS(u)}{du}$.

1.2.1 Función de supervivencia empírica

El método más simple para estimar la función de supervivencia es a través de la *función de supervivencia empírica* que, efectivamente, cuenta el número de datos mayores que t.

Definición 1.3. Dada una muestra aleatoria $T_1, T_2, T_3, \ldots, T_n$ con función de distribución T, la *función de supervivencia empírica* S_n se define, para todos los valores de t, como

$$S_n(t) = \frac{\text{Número de observaciones} \geq t}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{[t,\infty)}(T_i), \qquad (1.3)$$

y es un estimador de la función de supervivencia $S(t) = P(T \geq t)$.

1.2.2 La función de riesgo

La función de supervivencia examina la posibilidad de que ocurran muertes más allá de un punto dado en el tiempo. Para monitorear el tiempo de vida de un componente a través del soporte de la distribución del tiempo de vida se utiliza la *función de riesgo.*

La probabilidad de que un individuo muera en $(t, t+\Delta t)$, dado que ha sobrevivido hasta el tiempo t, está dada por

$$P(t \leq T < t+\Delta t \mid T \geq t) = \frac{P(t \leq T < t+\Delta t, T \geq t)}{P(T \geq t)} = \frac{P(t \leq T < t+\Delta t)}{P(T \geq t)}$$

Promediando sobre la longitud del intervalo de tiempo Δt se tiene la *tasa promedio de muerte* sobre el intervalo $(t, t+\Delta t)$. En el límite, cuando Δt tiende a cero, la tasa promedio de muerte se vuelve una tasa instantánea de muerte

$$\lim_{\Delta t \to \infty} \frac{P(t \leq T < t+\Delta t)}{(\Delta t)P(T \geq t)} = \lim_{\Delta t \to \infty} \frac{F(t+\Delta t) - F(t)}{(\Delta t)S(t)} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}$$

Esto motiva la siguiente definición.

Definición 1.5. Una variable aleatoria de supervivencia T tiene *función de riesgo, tasa de riesgo* o *fuerza de mortalidad*, h, definida para $t > 0$ como

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{\log S(t)\}.$$

Esta es la tasa instantánea de muerte al tiempo *t*, dado que el individuo ha sobrevivido hasta el tiempo *t*. Las funciones de riesgo registran cómo la tasa de muerte cambia con el tiempo. La función de supervivencia se expresa en términos de la función de riesgo de la siguiente manera:

$$S(t) = \exp\{-H(t)\}$$

donde

$$H(t) = \int_0^t h(u)\,du \qquad (1.4)$$

La función $H(t)$ es ampliamente utilizada en análisis de supervivencia y se le llama *riesgo acumulado* o *integrado*. El riesgo acumulado puede obtenerse a partir de la función de supervivencia

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

Ya que S_n es un estimador apropiado de *S*, se sigue que

$$H_n(t) = -\ln S_n(t)$$

es un estimador de $H(t)$. $H_n(t)$ es la llamada *función empírica acumulada de riesgo*.

1.2.3 Estimador Kaplan-Meier de la función de supervivencia

Para determinar el estimador Kaplan-Meier de la función de supervivencia de una muestra de datos de supervivencia censurados, se construye una serie de

intervalos de manera tal que sólo un tiempo de muerte está incluido en cada intervalo y es, a su vez, el inicio del intervalo.

Supóngase que hay n individuos con tiempos de supervivencia observados $t_1, t_2, t_3, \ldots, t_n$. Algunas de estas observaciones pueden estar censuradas por la derecha y puede haber varios individuos con el mismo tiempo de supervivencia. Supóngase ahora que hay r tiempos de muerte entre los individuos, donde $r \leq n$. Después de ordenar estos tiempos de muerte en orden ascendente, el j-ésimo se denota por $t_{(j)}$, para $j = 1, 2, \ldots, r$, así que los tiempos de muerte ordenados son $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. El número de individuos que están vivos justo antes del tiempo $t_{(j)}$ se denotará por n_j, para $j = 1, 2, \ldots, r$, y d_j denotará el número de los que mueren en ese momento. Entonces, el intervalo de tiempo $\left(t_{(j)} - \delta, t_{(j)}\right)$, donde δ es un intervalo de tiempo infinitesimal, incluye sólo un tiempo de muerte. Como hay n_j individuos que están vivos justo antes de $t_{(j)}$ y d_j muertes en $t_{(j)}$, la probabilidad de que un individuo muera durante el intervalo $\left(t_{(j)}, t_{(j)} + \delta\right)$ se estima por d_j / n_j. Entonces, la probabilidad de que un individuo sobreviva a través de ese intervalo es $\left(n_j - d_j\right)/n_j$.

Hay veces en que los tiempos de supervivencia censurados ocurren simultáneamente con alguna muerte; en este caso, al calcular los valores de las n_j, se considera al tiempo de supervivencia censurado como si hubiera ocurrido inmediatamente después del tiempo de muerte.

De la manera en que se construyeron los intervalos de tiempo, el intervalo $\left(t_{(j)}, t_{(j+1)} - \delta\right)$, no contiene muertes. Entonces, la probabilidad de sobrevivir a tal intervalo es uno y la probabilidad conjunta de sobrevivir a los intervalos $\left(t_{(j)} - \delta, t_{(j)}\right)$ y $\left(t_{(j)}, t_{(j+1)} - \delta\right)$ puede estimarse como $\left(n_j - d_j\right)/n_j$. En el límite,

conforme $\delta \to 0$, $\left(n_j - d_j\right)/n_j$ se vuelve un estimador de la probabilidad de sobrevivir al intervalo $\left(t_{(j)},\ t_{(j+1)}\right)$.

Ahora supóngase que las muertes de los individuos en la muestra ocurren independientemente una de otra. Entonces, la función de supervivencia estimada en cualquier momento en el *k*-ésimo intervalo $\left(t_{(k)},\ t_{(k+1)}\right)$, $k = 1, 2, \ldots, r$, donde $t_{(r+1)} := \infty$, será la probabilidad estimada de sobrevivir más allá de $t_{(k)}$, es decir, la probabilidad de sobrevivir al intervalo $\left(t_{(k)},\ t_{(k+1)}\right)$ y todos los anteriores. Este es el estimador Kaplan-Meier de la función de supervivencia, el cual está dado por

$$\hat{S}(t) = \prod_{j=1}^{k}\left(\frac{n_j - d_j}{n_j}\right) \tag{1.5}$$

para $t_{(k)} \le t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, con $\hat{S}(t) = 1$ para $t < t_{(1)}$ y donde $t_{(r+1)} := \infty$. Estrictamente hablando, si la observación más grande es un tiempo de supervivencia censurado t^*, $\hat{S}(t)$ está indefinido para $t > t^*$. Por otro lado, si el tiempo de supervivencia observado más grande, $t_{(r)}$, es una observación no censurada, $n_r = d_r$, y entonces $\hat{S}(t) = 0$ para $t \ge t_{(r)}$. Una gráfica del estimador Kaplan-Meier de la función de supervivencia es una función escalonada, en la cual las probabilidades de supervivencia estimadas son constantes entre tiempos de muerte adyacentes y decrece en cada tiempo de muerte.

El estimador Kaplan-Meier también es conocido como el *estimador de producto límite* de la función de supervivencia.

Nótese que si no hay tiempos de vida censurados en el conjunto de datos, $n_j - d_j = n_{j+1}$, $j = 1, 2, \ldots, r$, en la ecuación (1.5) y entonces se tiene que

$$\hat{S}(t) = \frac{n_{k+1}}{n_1},$$

para $k = 1, 2, \ldots, r-1$, con $\hat{S}(t) = 1$ para $t < t_{(1)}$ y $\hat{S}(t) = 0$ para $t \geq t_{(r)}$. Ahora, n_1 es el número de individuos en riesgo justo antes del primer tiempo de muerte, que es el número de individuos en la muestra, y n_{k+1} es el número de individuos con tiempos de supervivencia más grandes o iguales a t_{k+1}. Por consecuencia, en ausencia de censura, $\hat{S}(t)$ es simplemente la función de supervivencia empírica S_n definida en la ecuación (1.3).

1.2.4 Error estándar del estimador Kaplan-Meier

El estimador Kaplan-Meier de la función de supervivencia para algún valor t en el intervalo $\left(t_{(k)},\ t_{(k+1)}\right)$ puede escribirse como

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j ,$$

para $k = 1, 2, \ldots, r$, donde $\hat{p}_j = \left(n_j - d_j\right)/n_j$ es la probabilidad estimada de que un individuo sobreviva al intervalo que empieza en $t_{(j)}$, $j = 1, 2, \ldots, r$. Tomando logaritmos,

$$\log \hat{S}(t) = \sum_{j=1}^{k} \log \hat{p}_j ,$$

entonces la varianza de $\log \hat{S}(t)$ está dada por

$$\operatorname{var}\left\{\log \hat{S}(t)\right\} = \sum_{j=1}^{k} \operatorname{var}\left\{\log \hat{p}_j\right\} \qquad (1.6)$$

Ahora, puede suponerse que el número de individuos que sobreviven al intervalo que empieza en $t_{(j)}$ se distribuye binomial con parámetros n_j y p_j, donde p_j es la probabilidad real de sobrevivir al intervalo. El número observado de individuos que sobreviven es $n_j - d_j$, entonces

$$\text{var}\left(n_j - d_j\right) = n_j p_j \left(1 - p_j\right).$$

Como $\hat{p}_j = \left(n_j - d_j\right)/n_j$, entonces

$$\begin{aligned} \text{var}\left(\hat{p}_j\right) &= \text{var}\left(\frac{n_j - d_j}{n_j}\right) \\ &= \tfrac{1}{n_j^2}\,\text{var}\left(n_j - d_j\right) \\ &= \frac{p_j\left(1 - p_j\right)}{n_j}. \end{aligned}$$

La varianza de $\hat{p}_j$ puede estimarse entonces por

$$\frac{\hat{p}_j\left(1 - \hat{p}_j\right)}{n_j} \tag{1.7}$$

Para obtener la varianza de $\log\left(\hat{p}_j\right)$, se utiliza la aproximación por series de Taylor a la varianza de una variable aleatoria, dada por

$$\text{var}\left\{g(X)\right\} \approx \left\{\frac{\text{d}g(X)}{\text{d}X}\right\}^2 \text{var}(X) \tag{1.8}$$

Utilizando las expresiones (1.7) y (1.8) se tiene que

$$\begin{aligned}\operatorname{var}\left\{\log\left(\hat{p}_j\right)\right\} &\approx \frac{1}{\hat{p}_j^2}\operatorname{var}\left(\hat{p}_j\right)\\ &= \frac{\left(1-\hat{p}_j\right)}{\hat{p}_j n_j}\\ &= \frac{d_j}{n_j\left(n_j-d_j\right)}\end{aligned}$$

De la ecuación (1.6),

$$\operatorname{var}\left\{\log\hat{S}(t)\right\} \approx \sum_{j=1}^{k}\frac{d_j}{n_j\left(n_j-d_j\right)} \tag{1.9}$$

y aplicando otra vez la ecuación (1.8) se tiene que

$$\operatorname{var}\left(\log\hat{S}(t)\right) \approx \frac{1}{\left[\hat{S}(t)\right]^2}\operatorname{var}\left(\hat{S}(t)\right)$$

así que

$$\operatorname{var}\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]^2\sum_{j=1}^{k}\frac{d_j}{n_j\left(n_j-d_j\right)} \tag{1.10}$$

Finalmente, el error estándar el estimador Kaplan-Meier de la función de supervivencia está dada por

$$s.e.\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]\left\{\sum_{j=1}^{k}\frac{d_j}{n_j\left(n_j-d_j\right)}\right\}^{\frac{1}{2}} \tag{1.11}$$

para $t_{(k)} \le t < t_{(k+1)}$. Este resultado se conoce como *fórmula de Greenwood.*

Si no hay tiempos de supervivencia censurados, $n_j - d_j = n_{j+1}$, entonces

$$\sum_{j=1}^{k}\frac{d_j}{n_j\left(n_j-d_j\right)}=\sum_{j=1}^{k}\frac{n_j-n_{j+1}}{n_jn_{j+1}}$$
$$=\sum_{j=1}^{k}\left(\frac{1}{n_{j+1}}-\frac{1}{n_j}\right)$$
$$=\frac{n_1-n_{k+1}}{n_1n_{k+1}}$$
$$=\frac{1-\hat{S}(t)}{n_1\hat{S}(t)},$$

dado que $\hat{S}(t)=n_{k+1}/n_1$ para $t_{(k)}\le t<t_{(k+1)}$, $k=1,2,\ldots,r-1$, en la ausencia de censura. De ahí y de la ecuación (1.10), la varianza estimada de $\hat{S}(t)$ es $\hat{S}(t)\left[1-\hat{S}(t)\right]/n_1$. Este es un estimador de la varianza de la función de supervivencia empírica, S_n, bajo la suposición de que el número de individuos en riesgo al tiempo t tiene distribución binomial con parámetros $n_1, S(t)$.

1.2.5 Intervalos de confianza para los valores de la función de supervivencia.

Un intervalo de confianza para el valor real de la función de supervivencia al tiempo t se obtiene suponiendo que el valor estimado de la función de supervivencia en t se distribuye asintóticamente Normal con media $S(t)$ y varianza estimada dada por la ecuación (1.10). Así, un intervalo del $100(1-\alpha)\%$ de confianza para $S(t)$, para un valor t dado, es

$$\left(\hat{S}(t)-z_{1-\frac{\alpha}{2}}\text{s.e.}\left\{\hat{S}(t)\right\},\hat{S}(t)+z_{1-\frac{\alpha}{2}}\text{s.e.}\left\{\hat{S}(t)\right\}\right)$$

donde $\text{s.e.}\left\{\hat{S}(t)\right\}$ se obtiene de la ecuación (1.11) y $z_{1-\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$ de la distribución normal estándar.

Una dificultad con este procedimiento surge del hecho de que los intervalos de confianza son simétricos. Cuando la función de supervivencia estimada es cercana a cero o a uno, los intervalos simétricos son inapropiados, ya que pueden llevar a límites de confianza para la función de supervivencia que estarían fuera del intervalo $(0,1)$. Una solución pragmática a este problema es reemplazar cualquier límite mayor que la unidad por 1.0 y cualquier límite menor que cero por 0.0.

Un procedimiento alternativo es transformar $\hat{S}(t)$ de manera que tome valores en el rango $(-\infty,\infty)$ y obtener los intervalos de confianza para el valor transformado. Los límites de confianza resultantes se transformarán inversamente para obtener así intervalos de confianza para $S(t)$. Las transformaciones posibles son la logística, $\log\left[S(t)/\{1-S(t)\}\right]$, y la transformación log-log complementaria, $\log\{-\log S(t)\}$. Nótese que esta última cantidad es el logaritmo de la función de riesgo acumulada. En este caso, el error estándar del valor transformado de $\hat{S}(t)$ puede encontrarse utilizando la aproximación en la ecuación (1.8).

Otro problema es que en las colas de la distribución de los tiempos de supervivencia, es decir, cuando $\hat{S}(t)$ es cercano a cero o a uno, la varianza de $\hat{S}(t)$ obtenida utilizando la fórmula de Greenwood subestima el verdadero valor de la varianza. En estas circunstancias puede utilizarse una expresión alternativa para el error estándar de $\hat{S}(t)$. Peto et. al. (1977) propone que el error estándar de $\hat{S}(t)$ se obtenga por la ecuación

$$s.e.\{\hat{S}(t)\} = \frac{\hat{S}(t)\sqrt{1-\hat{S}(t)}}{\sqrt{n_k}},$$

para $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, donde n_k es el número de individuos en riesgo en $t_{(k)}$, el inicio del k-ésimo intervalo de tiempo construido.

Los errores estándar de $\hat{S}(t)$ obtenidos con esta fórmula tenderán a ser mayores de lo que deberían ser. Por esta razón, se recomienda utilizar el estimador de Greenwood para uso general.

1.3 COMPARACIÓN DE DOS GRUPOS DE DATOS DE SUPERVIVENCIA

En la comparación de dos grupos de datos de supervivencia hay métodos que pueden utilizarse para cuantificar hasta qué punto hay diferencias entre ellos. Se considerará ahora un procedimiento no paramétrico llamado *prueba de log-rangos (log-rank test).*

1.3.1 Prueba de log-rangos (log-rank test).

Se denotará a los dos grupos como Grupo I y Grupo II. Supóngase que hay r tiempos de muerte distintos, $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, a través de los dos grupos, y que al tiempo $t_{(j)}$ mueren d_{1j} individuos en el Grupo I y d_{2j} individuos en el Grupo II, para $j = 1, 2, \ldots, r$. A menos que dos o más individuos en un grupo tengan registrado el mismo tiempo de muerte, los valores de d_{1j} y d_{2j} serán cero o uno. Supóngase además que, justo antes del tiempo $t_{(j)}$, hay n_{1j} individuos en riesgo de muerte en el primer grupo y n_{2j} en el segundo grupo. Por consecuencia, al tiempo $t_{(j)}$ hay $d_j = d_{1j} + d_{2j}$ muertes en total de los $n_j = n_{1j} + n_{2j}$ individuos en riesgo.

Grupo	Número de muertes en $t_{(j)}$	Número de individuos que sobreviven más allá de $t_{(j)}$	Número de individuos en riesgo justo antes de $t_{(j)}$
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Tabla 1.1 Número de muertes en el j-ésimo tiempo de muerte en cada uno de los grupos de individuos.

Ahora considérese la hipótesis nula de que no hay diferencias en las experiencias de supervivencia de los individuos entre los dos grupos. Si los totales marginales en la Tabla 1.1 se consideran fijos, las cuatro entradas en esta tabla sólo se determinan por el valor de d_{1j}, el número de muertes en $t_{(j)}$ en el Grupo I. Puede considerarse entonces a d_{1j} como una variable aleatoria que tomará cualquier valor en el rango $\left(0, \min\{d_j, n_{1j}\}\right)$ y con distribución Hipergeométrica. De acuerdo con esto la probabilidad de que la variable aleatoria asociada con el número de muertes en el primer grupo tome el valor d_{1j} es

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}} \qquad (1.12)$$

La media de la variable hipergeométrica d_{1j} está dada por

$$e_{1j} = n_{1j} d_j / n_j \qquad (1.13)$$

así que e_{1j} es el número esperado de individuos que mueren al tiempo $t_{(j)}$ en el Grupo I.

Ahora, la diferencia entre el número total observado y el número esperado de muertes en el Grupo I es

$$U_L = \sum_{j=1}^{r} \left(d_{1j} - e_{1j} \right) \tag{1.14}$$

Esta estadística tendrá media cero, ya que $E\left(d_{1j}\right) = e_{1j}$. Más aún, como los tiempos de muerte son independientes unos de otros, la varianza de U_L es simplemente la suma de las varianzas de las d_{1j}. La varianza de d_{1j} está dada por

$$v_{1j} = \frac{n_{1j} n_{2j} d_j \left(n_j - d_j \right)}{n_j^2 \left(n_j - 1 \right)} \tag{1.15}$$

así que la varianza de U_L es

$$\text{var}\left(U_L\right) = \sum_{j=1}^{r} v_{1j} = V_L \tag{1.16}$$

Cuando el número de tiempos de muerte no es muy pequeño, por el Teorema Central del Límite, $U_L \sim N\left(0, V_L\right)$. Se sigue entonces que

$$\frac{U_L}{\sqrt{V_L}} \sim N\left(0,1\right),$$

y por consecuencia

$$\frac{U_L^2}{V_L} \sim \chi_1^2 \tag{1.17}$$

La prueba basada en esta estadística tiene varios nombres, incluyendo *Mantel-Cox* y *Peto-Mantel-Haenszel*, pero probablemente se le conoce más como prueba de log-rangos (log-rank test).

La estadística $W_L = U_L^2/V_L$ resume qué tanto se desvían los tiempos de supervivencia observados en los dos grupos de datos de los esperados, bajo la hipótesis nula de que no hay diferencia entre grupos. Mientras más grande sea el valor de esta estadística, es decir, mayor al cuantil $1-\frac{\alpha}{2}$ de la distribución ji-cuadrada con un grado de libertad, se tendrá más evidencia en contra de la hipótesis nula, con un nivel del $100(1-\alpha)\%$ de confianza.

1.4 Comparación de tres o más grupos de datos de supervivencia

La prueba de rangos puede extenderse al caso en el que se comparan tres o más grupos de datos de supervivencia. Supóngase que se van a comparar las distribuciones de supervivencia de g grupos de datos, para $g \geq 2$. Como una extensión de la estadística U_L se tiene

$$U_{Lk} = \sum_{j=1}^{r}\left(d_{kj} - \frac{n_{kj}d_j}{n_j}\right),$$

para $k = 1, 2, \ldots, g-1$. Estas cantidades se expresan entonces en forma de vector con $(g-1)$ entradas.

Se necesitan también las expresiones de las varianzas de U_{Lk} y para la covarianza entre pares de valores. En particular, la covarianza entre U_{Lk} y U'_{Lk} está dada por

$$V_{Lkk'} = \sum_{j=1}^{r} \frac{n_{kj} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right),$$

para $k, k' = 1, 2, \ldots, g-1$, donde $\delta_{kk'}$ es tal que

$$\delta_{kk'} = \begin{cases} 1 & \text{si } k = k' \\ 0 & \text{en otro caso.} \end{cases}$$

Estos términos forman la *matriz de varianzas-covarianzas*, V_L, que es simétrica y tiene las varianzas de las U_{Lk} en la diagonal y los términos de covarianza fuera de la diagonal.

Como $U_L \sim N_{g-1}(0, V_L)$ y, por consiguiente, $V_L^{-\frac{1}{2}} U_L \sim N_{g-1}(0, I)$, cuando la hipótesis nula es verdadera, la estadística $U_L' V_L^{-1} U_L$ se distribuye ji-cuadrada con $(g-1)$ grados de libertad. Para probar la hipótesis nula de que no hay diferencias entre grupos, se compara el valor de esta estadística con el cuantil $1 - \frac{\alpha}{2}$ de la distribución ji-cuadrada con $(g-1)$ grados de libertad; si es más grande que el cuantil se tendrá más evidencia en contra de la hipótesis nula, con un nivel del $100(1-\alpha)\%$ de confianza.

1.5 MODELANDO LOS DATOS DE SUPERVIVENCIA

En la mayoría de los estudios médicos que dan lugar a datos de supervivencia se tiene información suplementaria para cada individuo. Un ejemplo típico sería un ensayo clínico cuyo objetivo es comparar los tiempos de supervivencia de pacientes que reciben uno de 2 tratamientos. En un estudio como éste, variables demográficas como edad o sexo del paciente, valores de variables fisiológicas como presión arterial, factores asociados con el estilo de vida

del paciente, como si fuma o no y hábitos alimenticios, tendrán impacto en el tiempo que el paciente sobrevive. Por lo tanto, los valores de estas variables, a las que se les llama *variables explicativas*, deberán ser registrados desde el principio del estudio.

1.5.1 Modelando la función de riesgo

Por medio del análisis de datos de supervivencia se puede explorar cómo la supervivencia de un grupo de pacientes depende de los valores de una o varias variables explicativas, que bien pueden cambiar en el tiempo o tener valores que se registraron para cada paciente en el tiempo origen.

En el análisis de datos de supervivencia el interés se centra en el riesgo de muerte en cualquier momento después del tiempo origen del estudio. Como una consecuencia, en el análisis de supervivencia la función de riesgo se modela directamente. Los modelos resultantes son, de alguna manera, diferentes en forma de los modelos lineales utilizados en análisis de regresión o en el análisis de datos de un diseño de experimentos, donde se modela la dependencia de la media de la variable respuesta, o alguna función de ella, con algunas variables explicativas. Sin embargo, muchos de los principios y procedimientos usados en modelación lineal se utilizan en la modelación de datos de supervivencia.

Hay dos razones básicas para modelar los datos de supervivencia. Un objetivo del proceso de modelación es determinar cuál combinación de variables explicativas afecta la forma de la función de riesgo. En particular, puede estudiarse tanto el efecto que el tratamiento tiene en el riesgo de muerte, como el punto hasta el cuál otras variables explicativas afectan la función de riesgo. Otra razón para modelar la función de riesgo es obtener un estimador de la función en sí para un individuo. Además, puede encontrarse un estimador de la función de supervivencia, gracias la relación que existe entre ésta y la función de riesgo. Esto puede llevar a la estimación de cantidades tales como el tiempo mediano de supervivencia, que será función de las variables de interés en el modelo y que podrá estimarse en pacientes actuales o futuros con valores particulares de esas

variables explicativas. El estimador resultante puede ser útil al idear un régimen o aconsejar al paciente acerca de su pronóstico.

El modelo básico para datos de supervivencia considerado en el presente trabajo es el *modelo de riesgos proporcionales*. Este modelo fue propuesto por Cox (1972) y también se le conoce como *modelo de regresión de Cox*. Aunque el modelo se basa en la suposición de riesgos proporcionales, es decir, en que el riesgo de muerte en cualquier momento para un individuo en un grupo es proporcional al riesgo en ese momento para un individuo similar en el otro grupo, no se supone ninguna distribución de probabilidad sobre los tiempos de supervivencia, por esta razón este modelo es referido como *modelo semiparamétrico*.

1.5.1.1 Un modelo para comparar 2 grupos

Supóngase que los pacientes reciben aleatoriamente uno de dos tratamientos, que por lo general son el estándar y uno nuevo; y sean $h_E(t)$ y $h_N(t)$ los riesgos de muerte al tiempo t para los pacientes en el tratamiento estándar o nuevo, respectivamente. De acuerdo con un modelo simple para los tiempos de supervivencia de los dos grupos de pacientes, el riesgo al tiempo t para un paciente en el tratamiento nuevo es proporcional al riesgo en el mismo tiempo para un paciente en el tratamiento estándar. Este *modelo de riesgos proporcionales* puede expresarse de la siguiente manera

$$h_N(t) = \psi h_S(t) \tag{1.18}$$

para algún valor no negativo de t, donde ψ es una constante. Una implicación de esta suposición es que las correspondientes funciones de supervivencia para los individuos en diferentes tratamientos no se cruzan.

El valor de ψ es la razón entre el riesgo de muerte en cualquier momento para un individuo en el nuevo tratamiento y el de un individuo en el tratamiento

estándar, así que ψ se conoce como *riesgo relativo* o *razón de riesgos*. Si $\psi < 1$, el riesgo de muerte en t es menor para un individuo que toma el nuevo tratamiento, en relación al individuo que toma el estándar. El nuevo tratamiento es, entonces, mejor que el estándar. Por otro lado, si $\psi > 1$, el riesgo de muerte en t es mayor para un individuo en el nuevo tratamiento y, por lo tanto, el tratamiento estándar es superior.

Una manera alternativa de expresar al modelo en la ecuación (1.18) lleva a un modelo que puede generalizarse más fácilmente. Supóngase que se tienen datos de supervivencia para n individuos y sea $h_i(t)$, $i = 1, 2, \ldots, n$ la función de riesgo para el i-ésimo individuo. Sea $h_0(t)$ la función de riesgo para un individuo en el tratamiento estándar. La función de riesgo para un individuo en el nuevo tratamiento es, entonces, $\psi h_0(t)$. El riesgo relativo ψ no puede ser negativo así que es conveniente fijarlo como $\psi = \exp(\beta)$. El parámetro β es el logaritmo de la razón de riesgos, es decir, $\beta = \log(\psi)$ y cualquier valor real de β dará un valor positivo de ψ. Nótese que los valores positivos de β se obtienen cuando la razón de riesgos es mayor que uno, es decir, cuando el tratamiento nuevo es inferior al estándar.

Ahora, sea X una *variable indicadora* que toma el valor 0 si un individuo toma el tratamiento estándar y 1 si el individuo toma el nuevo. Si x_i es el valor que toma la variable X para el i-ésimo individuo en el estudio, $i = 1, 2, \ldots, n$, la función de riesgo para este individuo puede escribirse como

$$h_i(t) = e^{\beta x_i} h_0(t) \qquad (1.19)$$

donde $x_i = 1$ si el i-ésimo individuo pertenece el nuevo tratamiento y $x_i = 0$ en otro caso. Y de ahí

$$\beta = \log \frac{h_i(t)}{h_0(t)}.$$

Este es el *modelo de riesgos proporcionales* para la comparación entre dos grupos de tratamiento.

1.5.1.2 El modelo general de riesgos proporcionales

Ahora se generalizará el modelo anterior a la situación donde el riesgo de muerte en un momento en particular depende de los valores $x_1, x_2, \ldots, x_p$ de p variables explicativas $X_1, X_2, \ldots, X_p$. Se supone que los valores de estas variables han sido recolectados en el tiempo origen del estudio.

El conjunto de valores de las variables explicativas en el modelo de riesgos proporcionales será representado por un vector $\underline{x}$, tal que $\underline{x} = (x_1, x_2, \ldots, x_p)'$. Sea $h_0(t)$ la función de riesgo para un individuo cuyos valores de todas las variables explicativas que conforman el vector $\underline{x}$ son cero. A la función $h_0(t)$ se le llama *función de riesgo base*. La función de riesgo para el i-ésimo individuo puede entonces escribirse como

$$h_i(t) = \psi(\underline{x}_i) h_0(t),$$

donde $\psi(\underline{x}_i)$ es una función de los valores del vector de variables explicativas para el i-ésimo individuo. La función $\psi(\underline{x}_i)$ puede interpretarse como el riesgo al tiempo t para un individuo cuyo vector de variables explicativas es $\underline{x}_i$, con respecto al riesgo para un individuo para el cual $\underline{x} = 0$.

Una vez más, como el riesgo relativo $\psi(\underline{x}_i)$ no puede ser negativo, es conveniente escribirlo como $\exp(\eta_i)$, donde η_i es una combinación lineal de las p variables explicativas en $\underline{x}_i$. Por lo tanto,

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}.$$

En notación matricial $\eta_i = \underline{\beta}'\underline{x}_i$, donde $\underline{\beta}$ es el vector de coeficientes de las variables explicativas $x_1, x_2, \ldots, x_p$ en el modelo. Puede considerarse a β_j como el cambio en el logaritmo del cociente de riesgos cuando la variable x_{ji} aumenta su valor en uno y las demás variables explicativas se mantienen constantes, o a $\exp(\beta_j)$ como el cambio en el cociente de riesgos cuando la variable x_{ji} aumenta su valor en uno y las demás variables explicativas se mantienen constantes. A la cantidad η_i se le llama *componente lineal* del modelo, pero también se le conoce como *score de riesgo* para el *i*-ésimo individuo. El modelo de riesgos proporcionales entonces se vuelve

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) h_0(t) \qquad (1.20)$$

Como este modelo puede re-expresarse de la forma

$$\log\left(\frac{h_i(t)}{h_0(t)}\right) = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

el modelo de riesgos proporcionales puede ser considerado como un modelo lineal para el logaritmo del cociente de riesgos.

Nótese que no hay término constante en el componente lineal del modelo de riesgos proporcionales. Si un término constante β_0 fuera incluido, la función de riesgo base se reescalaría dividiendo $h_0(t)$ por $\exp(\beta_0)$, y el término constante se cancelaría. Como puede notarse no se han hecho suposiciones respecto a la forma real de la función de riesgo base $h_0(t)$.

En general, una función de riesgo puede depender de variables *continuas* y/o *discretas (o factores)*. Una variable continua es la que toma valores numéricos que frecuentemente están medidos en una escala continua, como la edad o la presión sistólica. Un factor es una variable que toma un número

limitado de valores, que se conocen como *niveles* del factor, por ejemplo, sexo, estado civil. Cada uno de estos dos tipos de variables se incorpora de manera diferente en el componente lineal de un modelo de riesgos proporcionales. En el caso de las variables discretas o factores es necesario parametrizar el modelo, ya que por identificabilidad no es posible ajustar un parámetro para cada nivel. La forma más común de hacerlo es elegir un nivel base.

1.5.2 Ajuste del modelo de riesgos proporcionales

Ajustar el modelo de riesgos proporcionales dado en la ecuación (1.20) a un conjunto observado de datos de supervivencia implica estimar los coeficientes desconocidos, $\beta_1, \beta_2, \ldots, \beta_p$, de las variables explicativas en el componente lineal del modelo, $X_1, X_2, \ldots, X_p$. También debe estimarse la función de riesgo base, $h_0(t)$. Ambos componentes del modelo deben estimarse de manera separada. Las β's se estiman primero utilizando el método de *máxima verosimilitud* y tales estimadores se utilizan después para construir el estimador de la función de riesgo base. Este resultado es importante, ya que significa que para hacer inferencias acerca de los efectos de las p variables explicativas $X_1, X_2, \ldots, X_p$ en el correspondiente riesgo, $h_i(t)/h_0(t)$, no se necesita estimar $h_0(t)$.

Supóngase que se tienen datos disponibles para n individuos, entre los cuales hay r tiempos de muerte distintos y $n-r$ tiempos de supervivencia censurados por la derecha. Supóngase también que sólo muere un individuo en cada tiempo de muerte, es decir, no hay *empates* en los datos. Los r tiempos de muerte son $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, de manera que $t_{(j)}$ es el j-ésimo tiempo de muerte ordenado. El conjunto de individuos en riesgo al tiempo $t_{(j)}$ será denotado por $R(t_{(j)})$, es decir, $R(t_{(j)})$ es el conjunto de individuos vivos y no censurados justo en el momento anterior a $t_{(j)}$. A la cantidad $R(t_{(j)})$ se le llama *conjunto de riesgo*.

La base del argumento utilizado en la construcción de la función de verosimilitud para el modelo de riesgos proporcionales es que los intervalos entre tiempos de muerte sucesivos no transmiten información acerca del efecto de las variables explicativas en el riesgo de muerte. Esto es porque la función de riesgo base tiene una forma arbitraria y, por lo tanto, es concebible que $h_0(t)$, y de ahí, $h(t)$, sea cero en aquellos intervalos de tiempo en los cuales no hay muertes. Esto significa que esos intervalos no proporcionan información acerca de los valores de los parámetros β. Por lo tanto, considérese la probabilidad de que el i-ésimo individuo muera en algún tiempo $t_{(j)}$, condicional en que uno muera en $t_{(j)}$, uno de los r tiempos de muerte observados $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$. Si el vector de variables explicativas para el individuo que muere en $t_{(j)}$ es $\underset{\sim}{x}_{(j)}$, esto es

$$P\left[\text{individuo con variables } \underset{\sim}{x}_{(j)} \text{ muere en } t_{(j)} \mid \text{una muerte en } t_{(j)}\right] \quad (1.21)$$

Por definición, esta expresión se convierte en

$$\frac{P\left[\text{individuo con variables } \underset{\sim}{x}_{(j)} \text{ muere en } t_{(j)}\right]}{P\left[\text{una muerte en } t_{(j)}\right]}.$$

El numerador de la expresión anterior es simplemente el riesgo de muerte en el tiempo $t_{(j)}$ para el individuo cuyo vector de variables explicativas es $\underset{\sim}{x}_{(j)}$. Si el i-ésimo individuo es el que muere en $t_{(j)}$, esta función de riesgo puede escribirse como $h_i\left(t_{(j)}\right)$. El denominador es la suma de los riesgos de muerte al tiempo $t_{(j)}$ sobre todos los individuos en el conjunto de riesgo en ese tiempo. La probabilidad condicional de la ecuación (1.21) queda expresada entonces como

$$\frac{h_i\left(t_{(j)}\right)}{\sum_{l \in R\left(t_{(j)}\right)} h_l\left(t_{(j)}\right)},$$

y, utilizando la ecuación (1.20), queda

$$\frac{\exp\left(\underline{\beta}'\underline{x}_{(j)}\right)}{\sum_{l\in R(t_{(j)})}\exp\left(\underline{\beta}'\underline{x}_l\right)}$$

Finalmente, tomando el producto de estas probabilidades condicionales sobre los r tiempos de muerte, la función de verosimilitud es

$$L\left(\underline{\beta}\right)=\prod_{j=1}^{r}\frac{\exp\left(\underline{\beta}'\underline{x}_{(j)}\right)}{\sum_{l\in R(t_{(j)})}\exp\left(\underline{\beta}'\underline{x}_l\right)} \qquad (1.22)$$

La función de verosimilitud que se ha obtenido no es una verosimilitud real, ya que no utiliza directamente los verdaderos tiempos de supervivencia censurados y no censurados. Por esta razón esta función es referida como *función de verosimilitud parcial.*

Los individuos con tiempos de supervivencia censurados no contribuyen al numerador de la función de log-verosimilitud, pero entran en la suma sobre los conjuntos de riesgo de los tiempos de muerte que ocurren antes del tiempo de censura. Más aún, la función de verosimilitud depende solamente del orden de los tiempos de muerte, ya que esto determina el conjunto de riesgo en cada uno de estos tiempos. Por consecuencia, las inferencias acerca del efecto de las variables explicativas en la función de riesgo dependen solamente en el orden de los tiempos de supervivencia.

Ahora supóngase que los datos consisten de n tiempos de supervivencia observados, denotados por $t_1, t_2, \ldots, t_n$ y que δ_i es un indicador de censura que toma el valor 0 si el i-ésimo tiempo de supervivencia t_i, $i=1,2,\ldots,n$, está censurado y 1 en otro caso. La función de verosimilitud de la ecuación (1.22) puede expresarse de la forma

$$\prod_{i=1}^{n}\left[\frac{\exp\left(\underline{\beta}'\underline{x}_i\right)}{\sum_{l\in R(t_i)}\exp\left(\underline{\beta}'\underline{x}_l\right)}\right]^{\delta_i},$$

donde $R(t_i)$ es el conjunto de riesgo al tiempo t_i. La correspondiente función de log-verosimilitud está dada por

$$\log L\left(\underline{\beta}\right)=\sum_{i=1}^{n}\delta_i\left\{\underline{\beta}'\underline{x}_i-\log\sum_{l\in R(t_i)}\exp\left(\underline{\beta}'\underline{x}_l\right)\right\} \qquad (1.23)$$

Los estimadores de máxima verosimilitud para los parámetros β en el modelo de riesgos proporcionales pueden encontrarse maximizando la función de log-verosimilitud utilizando métodos numéricos. Esta maximización generalmente se realiza utilizando el *procedimiento de Newton-Raphson.*

1.5.2.1 El procedimiento de Newton-Raphson

Sea $\underline{u}\left(\underline{\beta}\right)$ el vector de $p\times 1$, donde cada entrada es la primera derivada de la función de log-verosimilitud en la ecuación (1.23) con respecto al parámetro β_j, $j=1, 2, \ldots, p$. Esta cantidad se conoce como el *vector de scores de eficiencia.* También, sea la matriz de $p\times p$, $I\left(\underline{\beta}\right)$, la matriz de los negativos de las segundas derivadas de la log-verosimilitud, así que el (j,k)-ésimo elemento de $I\left(\underline{\beta}\right)$ es

$$-\frac{\partial^2\log L\left(\underline{\beta}\right)}{\partial\beta_j\partial\beta_k}$$

La matriz $I\left(\underline{\beta}\right)$ es conocida como la *matriz de información observada.*

De acuerdo con el procedimiento Newton-Raphson, un estimador del vector

de los parámetros, β, en el $(s+1)$-ésimo ciclo del procedimiento iterativo, $\underline{\hat{\beta}}_{s+1}$, es

$$\underline{\hat{\beta}}_{s+1} = \underline{\hat{\beta}}_s + I^{-1}\left(\underline{\hat{\beta}}_s\right)\underline{u}\left(\underline{\hat{\beta}}_s\right),$$

para $s = 0,1,2,\ldots$, donde $\underline{u}\left(\underline{\hat{\beta}}_s\right)$ es el vector de los scores de eficiencia y $I^{-1}\left(\underline{\hat{\beta}}_s\right)$ es el inverso de la matriz de información, ambos evaluados en $\underline{\hat{\beta}}_s$. El proceso puede empezarse tomando $\underline{\hat{\beta}}_0 = \underline{0}$. El proceso se termina cuando el cambio en la log-verosimilitud es suficientemente pequeño o cuando el más grande de los cambios relativos en los valores de los estimadores de los parámetros es suficientemente pequeño.

Cuando el procedimiento iterativo ha convergido, la matriz de varianzas-covarianzas de los estimadores de los parámetros puede aproximarse por la inversa de la matriz de información, evaluada en $\underline{\hat{\beta}}$, esto es, $I^{-1}\left(\underline{\hat{\beta}}\right)$. Las raíces cuadradas de los elementos de la diagonal de esta matriz son los errores estándar de los valores estimados de $\beta_1, \beta_2, \ldots, \beta_p$.

1.5.3 Intervalos de confianza y pruebas de hipótesis para las β's

Un intervalo del $100(1-\alpha)\%$ de confianza para el parámetro β_i, bajo la suposición de que $\hat{\beta}_i \sim N\left(\beta_i, I^{-1}\left(\hat{\beta}_i\right)\right)$, es

$$\left(\hat{\beta}_i - z_{1-\frac{\alpha}{2}}\text{s.e.}\left(\hat{\beta}_i\right), \hat{\beta}_i + z_{1-\frac{\alpha}{2}}\text{s.e.}\left(\hat{\beta}_i\right)\right)$$

donde $\hat{\beta}_i$ es el estimador de β_i y $z_{1-\frac{\alpha}{2}}$ es el cuantil $1-\dfrac{\alpha}{2}$ de la distribución normal estándar.

Si un intervalo del $100(1-\alpha)\%$ de confianza para β_i no incluye al cero, se tiene evidencia de que el valor de tal parámetro es estadísticamente diferente de cero. Más específicamente, la hipótesis nula $\beta_i = 0$ puede probarse calculando el valor de la estadística $\hat{\beta}_i / s.e.(\hat{\beta}_i)$. El valor observado de esta estadística se compara con los cuantiles de la distribución normal estándar. Este procedimiento algunas veces es llamado *prueba de Wald.*

En general, los estimadores individuales $\hat{\beta}_1, \hat{\beta}_2, \ldots$ en un modelo de riesgos proporcionales no son independientes uno de otro. Esto significa que los resultados de pruebas de hipótesis separadas para los parámetros β en un modelo no son fáciles de interpretar. Por ejemplo, considérese la situación en que hay tres variables explicativas X_1, X_2, X_3. Si $\hat{\beta}_1$ y $\hat{\beta}_2$ no son significativamente diferentes de cero al compararse con sus errores estándar, no se puede concluir que sólo X_3 debe ser incluida en el modelo. Esto es debido a que el coeficiente de X_1, por ejemplo, puede cambiar cuando X_2 es excluida del modelo y viceversa. Esto puede suceder cuando X_1 y X_2 están correlacionadas.

Debido a la dificultad al interpretar los resultados de las pruebas concernientes a los valores de los coeficientes de las variables explicativas en un modelo, se requieren métodos alternativos para comparar diferentes modelos de riesgos proporcionales que sean más satisfactorios que las pruebas de Wald.

1.5.4 Comparación de modelos alternativos

En el análisis de datos de supervivencia se buscan modelos que representen la dependencia de la función de riesgo con una o más variables explicativas. Durante el proceso se ajustan diferentes modelos de riesgos proporcionales que dependen de diferentes conjuntos de variables explicativas y se hacen comparaciones entre ellos.

Como un ejemplo específico, considérese la situación donde hay dos grupos de tiempos de supervivencia, correspondientes a individuos que reciben un tratamiento estándar o uno nuevo. La función de riesgo común bajo el modelo para el cual no hay diferencia entre tratamientos puede tomarse como $h_0(t)$. Este modelo es un caso especial del modelo general de riesgos proporcionales de la ecuación (1.20) en el cual no hay variables explicativas en el componente lineal del modelo. A este modelo se le llama *modelo nulo.*

En la aproximación por modelos al análisis de datos de supervivencia hay dos pasos que deben realizarse. Primero, se necesita examinar el ajuste de un modelo de riesgos proporcionales a los datos observados para asegurar que es el apropiado. Segundo, se necesita interpretar el modelo para cuantificar el efecto que tienen las variables explicativas en la función de riesgo. Pero antes, deben considerarse ciertas estrategias para la selección de modelos.

1.5.5 Estrategias para la selección de modelos

Un paso inicial en el proceso de selección de modelos es identificar un conjunto de variables explicativas que tienen potencial para ser incluidas en el componente lineal de un modelo de riesgos proporcionales. Este conjunto debe contener las variables y factores que se han recolectado para cada individuo, pero adicionalmente se pueden requerir los términos que correspondan a las interacciones entre factores o entre variables y factores.

Una vez que se ha aislado el conjunto de variables explicativas potenciales, se debe determinar la combinación de variables que se utilizará al modelar la función de riesgo. En la práctica, una función de riesgo no dependerá de una combinación única de variables. Más bien se tendrá cierto número de buenos modelos, más que un único "mejor" modelo. Por esta razón es deseable considerar un rango más amplio de modelos posibles.

Un principio importante en la modelación estadística es que, cuando un

modelo tiene un término de interacción, los correspondientes términos de orden menor deben estar incluidos también. Esta regla se conoce como *principio jerárquico.*

La estrategia de selección de modelos depende hasta cierto punto del propósito del estudio. En algunas aplicaciones se obtendrá información de cierto número de variables y el objetivo será determinar cuáles de ellas tienen efecto en la función de riesgo. En otras situaciones habrá una o más variables de interés primario, tales como términos correspondientes al efecto de un tratamiento. El objetivo del proceso de modelación será evaluar el efecto de tales variables en la función de riesgo. Como tal vez se espere que las otras variables que se han registrado influyan en el tamaño y precisión del efecto del tratamiento, estas variables tendrán que ser tomadas en cuenta en el proceso de modelación.

1.5.5.1 Procedimientos para la selección de variables

Primero se considerará la situación en la que todas las variables tienen igual importancia y el objetivo es identificar subconjuntos de ellas de los cuales depende la función de riesgo. Cuando el número de variables explicativas potenciales, incluyendo interacciones, términos no lineales, etc. no es muy grande, puede ser factible ajustar todas las posibles combinaciones de términos, prestando la debida atención al principio jerárquico. Los modelos anidados alternativos pueden ser comparados examinando el cambio en el valor de $-2\log \hat{L}$ al añadir o quitar términos del modelo o bien por medio de la *devianza.*

La devianza es la estadística que se utiliza para resumir qué tanto el modelo que se está probando difiere de un modelo que ajusta perfectamente a los datos. A este último modelo se le llama saturado o completo, y es el modelo en el cual se permite que los coeficientes β sean diferentes para cada individuo. Esta estadística está dada por

$$D = -2\left\{\log \hat{L}_a - \log \hat{L}_s\right\}$$

donde $\hat{L}_a$ es la verosimilitud parcial maximizada bajo el modelo actual y $\hat{L}_s$ es la verosimilitud parcial maximizada bajo el modelo saturado. Mientras más pequeño sea el valor de la devianza, mejor será el modelo.

Las comparaciones entre modelos, que no necesitan estar anidados, pueden hacerse también en base a la estadística

$$AIC = -2\log \hat{L} + \alpha p ,$$

donde p es el número de parámetros desconocidos en el modelo y α es una constante predeterminada. El valor de α usualmente se toma entre 2 y 6, (aunque el más recomendado para uso general es 3 (Collet, D (1994)). Esta estadística se conoce como *Criterio de información de Akaike* y mientras más pequeño sea su valor, mejor será el modelo. La motivación detrás de esta estadística es que si la única diferencia entre dos modelos es que uno incluye covariables innecesarias, los valores de *AIC* para los dos modelos no serán muy diferentes. Es más, el valor de *AIC* tenderá a incrementarse cuando se agreguen términos innecesarios en el modelo. Es importante que después de elegir un modelo se confirme que éste ajusta correctamente a los datos utilizando los métodos de revisión de modelos, que se verán posteriormente.

Cuando el número de covariables es relativamente grande, el número de modelos posibles que se necesitará ajustar será computacionalmente caro. En esta situación se pueden utilizar rutinas basadas en los métodos *forward*, *backward*, o una combinación de ambos llamada *stepwise*.

Estas rutinas automáticas tienen cierto número de desventajas. Típicamente, llevan a la identificación de un subconjunto de variables en particular, en lugar de un conjunto de modelos igualmente buenos. Los subconjuntos encontrados por estas rutinas frecuentemente dependerán del procedimiento de selección que se utilizó, es decir, si fue selección hacia adelante (forward), hacia atrás (backward) o stepwise y generalmente tenderán a ignorar el

principio jerárquico. También dependen de una regla de paro que se utiliza para determinar si un término debe ser incluido o excluido de un modelo. Por todas estas razones, estas rutinas deben tener un papel limitado en la selección de modelos y deben ser utilizadas con cuidado.

1.6 REVISION DE MODELOS

Después de que se ha ajustado un modelo a un conjunto de datos de supervivencia observados se necesita evaluar si es adecuado. El uso de procedimientos de diagnóstico para la validación de modelos es una parte esencial del proceso de modelación.

En algunas situaciones, la inspección cuidadosa de los datos observados puede llevar a la identificación de ciertas características, tales como individuos con tiempos de supervivencia inusualmente grandes o pequeños. Sin embargo, a menos que haya solamente una o dos variables explicativas, una examinación visual de los datos no es muy útil. La situación se complica más cuando existe censura, ya que es más difícil juzgar si el modelo es adecuado o no, aún en las situaciones más sencillas. Es por esto que la examinación visual de los datos debe complementarse con procedimientos más formales de detección de insuficiencias en los modelos ajustados.

Una vez que un modelo ha sido ajustado, hay ciertos aspectos del ajuste que necesitan estudiarse. Por ejemplo, de todas las variables explicativas medidas durante el estudio, el modelo debe incluir un conjunto apropiado de estas. Por consiguiente, se necesitan procedimientos que permitan revisar si una variable incluida en el modelo necesita transformarse o si alguna variable omitida debería incluirse. También será importante identificar si hay tiempos de supervivencia que sean más grandes de lo que se pudo haber anticipado o individuos cuyas variables explicativas tienen un impacto excesivo en la razón de riesgos. Además, se necesita alguna forma de revisar la suposición de riesgos proporcionales al ajustar un modelo.

Muchos de los procedimientos de revisión de modelos están basados en los *residuos*, que se calculan para cada individuo en el estudio y tienen la característica de que su comportamiento es conocido, al menos aproximadamente, cuando el ajuste del modelo es satisfactorio.

1.6.1 Residuos para el modelo de regresión de Cox

Supóngase que se tienen los tiempos de supervivencia de n individuos, donde r son tiempos de muerte y los restantes $n-r$ están censurados por la derecha. Supóngase además que se ha ajustado el modelo de regresión de Cox a los tiempos de supervivencia y que el componente lineal del modelo contiene p variables explicativas $X_1, X_2, \ldots, X_p$. La función ajustada de riesgo para el i-ésimo individuo, $i = 1, 2, \ldots, n$, es entonces

$$\hat{h}_i(t) = \exp\left(\hat{\beta}'\underline{x}_i\right)\hat{h}_0(t),$$

donde $\hat{\beta}'\underline{x}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi}$ es el valor del predictor lineal ajustado del modelo para tal individuo y $\hat{h}_0(t)$ es la función estimada de riesgo base.

1.6.1.1 Residuos de Cox-Snell

Se le llama así porque es un ejemplo particular de la definición general de residuos dada por Cox y Snell (1968).

El residuo Cox-Snell para el i-ésimo individuo, $i = 1, 2, \ldots, n$, está dado por

$$r_{Ci} = \exp\left(\hat{\beta}'\underline{x}_i\right)\hat{H}_0(t_i) \qquad (1.24)$$

donde $\hat{H}_0(t_i)$ es la función acumulada estimada de riesgo base al tiempo t_i, el tiempo de supervivencia observado para tal individuo. Nótese que este residuo es

el valor de $\hat{H}_i(t_i) = -\log \hat{S}_i(t_i)$, donde $\hat{H}_i(t_i)$ y $\hat{S}_i(t_i)$ son los valores estimados del riesgo acumulado y de la función de supervivencia, respectivamente, para el *i*-ésimo individuo al tiempo t_i.

De acuerdo al teorema de cambio de variable, si T es la variable aleatoria asociada con el tiempo de supervivencia de un individuo y $S(t)$ es la correspondiente función de supervivencia, entonces la función de densidad de la variable aleatoria $Y = -\log S(t)$ está dada por

$$f_Y(y) = \left|\frac{dy}{dt}\right|^{-1} f_T\left\{S^{-1}\left(e^{-y}\right)\right\} \qquad (1.25)$$

donde $f_T(t)$ es la función de densidad de T. Ahora

$$\frac{dy}{dt} = \frac{d\{-\log S(t)\}}{dt} = \frac{f_T(t)}{S(t)},$$

y sustituyendo $t = S^{-1}\left(e^{-y}\right)$ en la igualdad anterior se tiene que

$$\frac{f_T\left\{S^{-1}\left(e^{-y}\right)\right\}}{S\left\{S^{-1}\left(e^{-y}\right)\right\}} = \frac{f_T\left\{S^{-1}\left(e^{-y}\right)\right\}}{e^{-y}}.$$

Finalmente, sustituyendo esta derivada en la ecuación (1.25), se tiene que

$$f_Y(y) = e^{-y},$$

por lo tanto, Y tiene una distribución exponencial con parámetro 1.

Si el modelo ajustado a los datos observados es satisfactorio, entonces un estimador, basado en el modelo, de la función de supervivencia para el *i*-ésimo

individuo al tiempo t_i, el tiempo de supervivencia para tal individuo, será cercano al correspondiente valor real $S_i(t_i)$. Esto sugiere que si se ha ajustado el modelo correcto, los valores $\hat{S}_i(t_i)$ tendrán propiedades similares a las de $S_i(t_i)$. Entonces, los negativos de los logaritmos de las funciones de supervivencia estimadas, $-\log \hat{S}_i(t_i)$, $i = 1, 2, \ldots, n$, serán una muestra de tamaño n de una distribución exponencial con parámetro unitario. Estos son los estimadores de los residuos de Cox-Snell.

Si el tiempo de supervivencia observado para un individuo está censurado por la derecha entonces el correspondiente valor del residuo también estará censurado por la derecha. Los residuos serán, entonces, una muestra censurada de una densidad exponencial con parámetro uno y al probar esta suposición se puede saber si el modelo es adecuado.

Los residuos de Cox-Snell r_{Ci} tienen varias propiedades. En particular, como tienen una distribución exponencial cuando se ajusta un modelo apropiado, no toman valores negativos; más aún, tendrán media y varianza uno. Otro punto a notar es que si el tiempo de supervivencia más grande no está censurado, el valor estimado de la función de supervivencia más allá de ese tiempo es cero, y el residuo r_{Ci} estará indefinido para tal observación.

1.6.1.2 Residuos de Cox-Snell modificados

Las observaciones censuradas llevan a residuos que no pueden ser considerados de la misma manera que los derivados de observaciones no censuradas. Por consiguiente, puede buscarse modificar los residuos de Cox-Snell de tal manera que se tome en cuenta explícitamente la censura.

Supóngase que el *i*-ésimo tiempo de supervivencia es una observación censurada, t_i^*, y sea t_i el tiempo de supervivencia real (desconocido), tal que

$t_i > t_i^*$. El residuo de Cox-Snell para esta individuo, evaluado en el tiempo de vida censurado, está dado por

$$r_{Ci} = \hat{H}_i\left(t_i^*\right) = -\log \hat{S}_i\left(t_i^*\right),$$

donde $\hat{H}_i\left(t_i^*\right)$ y $\hat{S}_i\left(t_i^*\right)$ son los valores estimados del riesgo acumulado y de la función de supervivencia, respectivamente, para el *i*-ésimo individuo al tiempo de supervivencia censurado.

Si el modelo ajustado es correcto, entonces se puede considerar que los valores r_{Ci} tienen una distribución exponencial con parámetro uno. La función de riesgo acumulada de esta distribución aumenta linealmente con el tiempo, así que mientras más grande es el valor del tiempo de supervivencia t_i para el *i*-ésimo individuo, más grande es el valor del residuo Cox-Snell para el individuo. Se sigue, entonces, que el residuo para el *i*-ésimo individuo en el tiempo de muerte real (desconocido), $\hat{H}_i\left(t_i\right)$, será más grande que el residuo evaluado en el tiempo de supervivencia censurado observado.

Para tomar en cuenta esto, los residuos de Cox-Snell pueden modificarse con la adición de una constante positiva Δ, que puede ser llamada *exceso del residuo (excess residual).* Los residuos de Cox-Snell modificados son, entonces, de la forma

$$r'_{Ci} = \begin{cases} r_{Ci} & \text{para observaciones no censuradas,} \\ r_{Ci} + \Delta & \text{para observaciones censuradas,} \end{cases}$$

donde r_{Ci} es el residuo de Cox-Snell para la *i*-ésima observación, definido en la ecuación (1.24). Ahora sólo queda identificar un valor satisfactorio para Δ. Para esto, se utiliza la propiedad de pérdida de memoria de la distribución exponencial. De acuerdo con este resultado, como r_{Ci} tiene una distribución exponencial con parámetro uno, el exceso del residuo Δ tendrá la misma

distribución. El valor esperado de Δ es uno, sugiriendo que Δ deba ser la unidad, lo cual lleva a los siguientes residuos de Cox-Snell modificados

$$r'_{Ci} = \begin{cases} r_{Ci} & \text{para observaciones no censuradas,} \\ r_{Ci} + 1 & \text{para observaciones censuradas.} \end{cases} \qquad (1.26)$$

El i-ésimo residuo de Cox-Snell modificado puede expresarse de una manera alternativa introduciendo un indicador de censura, δ_i, que toma el valor de cero si el tiempo de supervivencia observado para el i-ésimo individuo está censurado y uno si no lo está. Entonces, el residuo de Cox-Snell modificado está dado por

$$r'_{Ci} = 1 - \delta_i + r_{Ci} \qquad (1.27)$$

Nótese que de la definición de este tipo de residuo, r'_{Ci} debe ser más grande que uno para una observación censurada. Además, como para los residuos no modificados, r'_{Ci} puede tomar cualquier valor entre cero e infinito y tendrán una distribución sesgada.

Con base en evidencia empírica, Crowley y Hu (1977) encontraron que la suma de la unidad al residuo Cox-Snell para una observación censurada inflaba demasiado el residuo. Ellos sugirieron entonces que era mejor utilizar el valor de la mediana en lugar del de la media. Para la distribución exponencial con parámetro uno el valor de la mediana es $\ln 2 = 0.693$. Así, una segunda versión del residuo de Cox-Snell modificado es

$$r''_{Ci} = \begin{cases} r_{Ci} & \text{para observaciones no censuradas,} \\ r_{Ci} + \ln 2 & \text{para observaciones censuradas.} \end{cases} \qquad (1.28)$$

Cuando la proporción de observaciones censuradas no es muy grande, el conjunto de residuos obtenido con la primera de estas modificaciones no será muy diferente del obtenido con la segunda.

1.6.1.3 Residuos de Martingala

Los residuos modificados r'_{Ci} definidos en la ecuación (1.27) tienen media igual a uno para las observaciones no censuradas. De acuerdo con esto, esos residuos pueden ser refinados trasladando los r'_{Ci} de tal manera que tengan media cero cuando la observación está censurada. Si, además, los valores resultantes se multiplican por -1, se obtienen los siguientes residuos

$$r_{Mi} = \delta_i - r_{Ci} \tag{1.29}$$

Estos residuos se conocen como *residuos de martingala*, ya que pueden también derivarse utilizando métodos de martingalas (Flemming y Harrington (1991)).

Los residuos de martingala toman valores en el intervalo $(-\infty, 1)$, siendo negativos los residuos para las observaciones censuradas, donde $\delta_i = 0$. En muestras grandes los residuos de martingala no están correlacionados entre sí y tienen un valor esperado de cero; sin embargo, no están distribuidos simétricamente alrededor del cero.

Otra forma de ver los residuos de martingala es notar que la cantidad r_{Mi} de la ecuación (1.29) es la diferencia entre el número observado de muertes para el i-ésimo individuo en el intervalo $(0, t_i)$ y el correspondiente número esperado estimado en base al modelo ajustado. Para ver esto, nótese que el número observado de muertes es uno si el tiempo de supervivencia t_i no está censurado, y cero si lo está, esto es δ_i. El segundo término en la ecuación (1.29) es un estimador de $H_i(t_i)$, el riesgo acumulado o la probabilidad acumulada de muerte para el i-ésimo individuo sobre el intervalo $(0, t_i)$. Como estamos tratando con sólo un individuo, esto puede verse como el número esperado de muertes en tal intervalo.

1.6.1.4 Residuos de devianza

Como los residuos de martingala no están distribuidos simétricamente alrededor del cero, aun cuando el modelo ajustado es correcto, las gráficas basadas en estos residuos son difíciles de interpretar. Los residuos de devianza, que fueron introducidos por Therneau, Grambsch y Fleming (1990) están distribuidos más simétricamente alrededor del cero. Están definidos como

$$r_{Di} = \operatorname{sgn}(r_{Mi})\left[-2\left\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\right\}\right]^{\frac{1}{2}} \qquad (1.30)$$

donde r_{Mi} es el residuo de martingala para el *i*-ésimo individuo. La función $\operatorname{sgn}(r_{Mi})$ asegura que los residuos de devianza tengan el mismo signo que los residuos de martingala.

La motivación original para estos residuos es que son componentes de la devianza. Los residuos de devianza son tales que $D = \sum r_{Di}^2$, así que las observaciones que corresponden a residuos de devianza relativamente grandes son aquellas que no ajusta bien el modelo.

Otra forma de ver los residuos de devianza es que son los residuos de martingala transformados para producir valores que son simétricos alrededor del cero cuando el modelo ajustado es apropiado. Para entender mejor esto, primero hay que recordar que los residuos de martingala r_{Mi} pueden tomar cualquier valor en el intervalo $(-\infty, 1)$. Para valores negativos grandes de r_{Mi}, el término entre los corchetes en la expresión (1.30) está dominado por r_{Mi}. El tomar la raíz cuadrada de esta cantidad tiene el efecto de llevar el residuo más cerca de cero. Así que los residuos de martingala en el rango $(-\infty, 0)$ se comprimen contra el cero. Ahora hay que considerar los residuos de martingala en el intervalo $(0,1)$. El término $\delta_i \log(\delta_i - r_{Mi})$ en la ecuación (1.30) sólo será diferente de cero para observaciones censuradas y tomarán, entonces, el valor $\log(1 - r_{Mi})$. Conforme r_{Mi} se acerca a

uno, $1-r_{Mi}$ se acerca a cero y $\log(1-r_{Mi})$ toma valores negativos más grandes. La cantidad en corchetes en la expresión (1.30) es dominada entonces por este término logarítmico, de tal manera que los residuos de devianza se van acercando hacia $+\infty$ conforme los residuos de martingala se acercan a uno.

Como punto final hay que notar que aunque se espera que estos residuos se distribuyan simétricamente alrededor del cero cuando se ajusta un modelo apropiado, no necesariamente suman cero.

1.6.1.5 Residuos de Schoenfeld

Dos desventajas de los residuos descritos anteriormente es que dependen en gran manera del tiempo de supervivencia observado y requieren un estimador de la función de riesgo acumulada. Ambas desventajas son superadas por un residuo propuesto por Schoenfeld (1982) llamado *residuo de Schoenfeld*. Este residuo difiere de los considerados anteriormente en que no se tiene un valor único del residuo para cada individuo, sino un conjunto de valores, uno para cada variable explicativa incluida en el modelo de regresión de Cox ajustado.

El *i*-ésimo residuo de Schoenfeld para X_j, la j-ésima variable explicativa en el modelo, está dado por

$$r_{Sji} = \delta_i \left\{ x_{ji} - a_{ji} \right\} \tag{1.31}$$

donde δ_i es la variable indicadora antes descrita, x_{ji} es el valor de la j-ésima variable explicativa, $j = 1, 2, \ldots, p$, para el *i*-ésimo individuo en el estudio,

$$a_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp\left(\hat{\beta}\underline{x}_l\right)}{\sum_{l \in R(t_i)} \exp\left(\hat{\beta}\underline{x}_l\right)} \tag{1.32}$$

y $R(t_i)$ es el conjunto de todos los individuos en riesgo al tiempo t_i. Nótese que

los valores diferentes de cero de estos residuos sólo surgen de observaciones no censuradas.

Como los estimadores de las $\beta's$ son tales que

$$\left.\frac{\partial \log L(\underline{\hat{\beta}})}{\partial \beta_j}\right|_{\underline{\hat{\beta}}} = 0$$

los residuos de Schoenfeld deben sumar cero. Estos residuos también tienen la propiedad de que en muestras grandes el valor esperado de $r_{S\hat{\beta}i}$ es cero y no están correlacionados entre sí.

Al utilizar estos residuos hay que notar que los individuos menos probables de morir al tiempo t_i, con respecto a aquellos en riesgo de muerte en t_i, tendrán residuos con valores pequeños. Por el contrario, los individuos que tienen más probabilidad de morir, con respecto a aquellos en riesgo, tendrán residuos relativamente grandes.

1.6.2 Gráficas basadas en residuos

Como se supone que los residuos de Cox-Snell se comportan como observaciones de una distribución exponencial unitaria cuando el modelo ajustado es correcto, tienen media y varianza unitarias. Esto significa que las gráficas simples, tales como las gráficas de residuos contra el número de observación, conocidas como *gráficas de índice*, no serán simétricas. Por esta razón, las gráficas no son fáciles de interpretar y se necesita algo un poco más sofisticado.

Sea $\hat{S}(t)$ el estimador Kaplan-Meier de la función de supervivencia para los datos. Después de calcular los residuos de Cox-Snell, r_{Ci}, se encuentra el estimador de Kaplan-Meier para la función de supervivencia de estos valores.

Este estimador se calcula de manera similar que el estimador Kaplan-Meier para la función de supervivencia de los tiempos de supervivencia, excepto que los datos en los cuales se basa el estimador ahora son los residuos r_{Ci}. Los residuos obtenidos de observaciones censuradas se toman como censurados. Denotando el estimador por $\hat{S}(r_{Ci})$, se grafican los valores de $\log\{-\log\hat{S}(r_{Ci})\}$ contra $\log r_{Ci}$. Esto da una gráfica de riesgo log-acumulada para los residuos. Una gráfica con una línea recta con pendiente uno y con intersección en el cero indicará que el modelo ajustado es correcto. Por otro lado, una gráfica que muestre cierta lejanía de la línea recta o una recta que no tenga aproximadamente pendiente uno o intersección en el cero, indicará que el modelo necesita modificarse de alguna manera.

Cuando un modelo incluya factores, se harán gráficas de riesgo log-acumulativas de los residuos de Cox-Snell para cada nivel de cada factor en el modelo. Si el modelo ajustado es satisfactorio, los puntos en la gráfica correspondientes a los diferentes niveles deberán traslaparse. Por otro lado, si los residuos para un nivel del factor aparecen generalmente separados de aquellos en los otros niveles, esto sugiere que el factor no ha sido tomado en cuenta de manera apropiada en el modelo.

1.6.2.1 Gráficas basadas en otros tipos de residuos

Cuando se han ajustado los residuos de Cox-Snell en presencia de censura, pueden obtenerse gráficas basadas en los residuos modificados. Por ejemplo, se puede graficar los residuos de Cox-Snell modificados contra el número de observación para obtener una gráfica de índice. Los residuos con valores negativos relativamente grandes pueden indicar que la observación correspondiente no ha sido ajustada de manera adecuada en el modelo.

Se ha visto que lo que determina los valores de los parámetros β es el orden de los tiempos de supervivencia, más que los tiempos de supervivencia en sí. De acuerdo con esto, las gráficas de los residuos contra los tiempos de supervivencia

ordenados pueden utilizarse para examinar si el modelo es adecuado. Si en tales gráficas no se presenta ningún patrón el modelo ajustado es factible.

Los residuos de Cox-Snell modificados pueden graficarse también contra las variables explicativas que no fueron incluidas en el modelo de regresión de Cox. Si no existe ninguna relación obvia entre los residuos y alguna de tales variables, es indicativo de que la variable no se necesita en el modelo. De la misma manera, las gráficas de residuos contra las variables incluidas en el modelo pueden utilizarse como una comprobación informal de la necesidad de tales variables. Las gráficas de los residuos contra las variables explicativas también pueden utilizarse para indicar si alguna variable en particular necesita transformarse antes de incorporarla en el modelo.

Aunque se permite la censura en la construcción de los residuos de Cox-Snell modificados, es buena idea utilizar símbolos diferentes para las observaciones censuradas y las no censuradas al graficar. Esto ayuda a encontrar posibles características inusuales asociadas con la censura.

Los residuos de martingala pueden utilizarse en la construcción de las gráficas que se describieron para los residuos de Cox-Snell modificados. En particular, las gráficas de índice de los residuos, las gráficas contra los tiempos de supervivencia o los tiempos de supervivencia ordenados y las gráficas contra las variables explicativas dentro o fuera del modelo pueden ser útiles en la validación de modelos. Las gráficas de índice pueden revelar observaciones que no fueron correctamente ajustadas por el modelo, y las gráficas contra las variables explicativas indican si la variable necesita incluirse o si es necesario transformar alguna de las variables incluidas en el modelo. Las gráficas de residuos contra los tiempos de supervivencia ordenados pueden utilizarse para detectar desviaciones de los riesgos proporcionales.

También se pueden hacer las mismas gráficas para los residuos de devianza. Gracias a la simetría que tienen estos residuos cuando se ajusta el modelo correcto, las gráficas construidas serán más fáciles de interpretar. Por

ejemplo, una gráfica de índice resaltará las observaciones con valores positivos o negativos relativamente grandes, haciendo más fácil identificar cuáles observaciones no ajusta bien el modelo.

Cuando un modelo incluye uno o más factores, los puntos en una gráfica de índice para los residuos pueden etiquetarse de acuerdo a los niveles de un factor en particular. Debe observarse el mismo patrón para las observaciones censuradas y las no censuradas en cada nivel del factor.

Las gráficas de los residuos de Schoenfeld contra los tiempos de supervivencia deben mostrar una dispersión aleatoria de los puntos, centrados en cero, si el modelo ajustado es el adecuado. Si el residuo de Schoenfeld se calcula para una variable indicadora utilizada para ajustar un factor en el modelo, una gráfica de los residuos contra los tiempos de supervivencia ordenados tendrá un patrón determinado. Típicamente, la gráfica mostrará dos bandas horizontales de puntos correspondientes a cada valor de la variable, una de cada lado de la línea que representa al residuo con valor cero.

1.6.3 Algunos comentarios y recomendaciones

En una sección anterior se discutió que, como los valores $-\log S(t_i)$ tienen una distribución exponencial con parámetro uno, los residuos de Cox-Snell, que son estimadores de estas cantidades, deberían tener aproximadamente una distribución exponencial cuando el modelo ajustado es correcto. Este resultado se utilizó después cuando se interpretó la gráfica del riesgo log-acumulativo de los residuos. Desafortunadamente esta aproximación no es muy fiable, particularmente en muestras pequeñas. Esto sucede porque al calcular los residuos de Cox-Snell no se utilizan los estimadores de las β's ni de la función de riesgo acumulado, $H_0(t)$. La sustitución de estimadores significa que la distribución de los residuos no necesariamente es exponencial unitaria, pero su distribución exacta no se conoce.

Crowley y Storer mostraron empíricamente que una gráfica de riesgos log-acumulativos de los residuos no es particularmente buena para identificar las insuficiencias del modelo ajustado. Más aún, en el caso particular del modelo nulo, la gráfica de riesgo log-acumulativo será una línea recta con pendiente uno y con intersección en cero, aún cuando debería haber algunas variables explicativas incluidas en el modelo. La razón para que suceda esto es que, cuando no se incluye ninguna covariable, el residuo de Cox-Snell para el i-ésimo individuo se reduce a $-\log \hat{S}_0(t_i)$. En ausencia de empates, esto es, aproximadamente $\sum_{j=1}^{k} 1/n_j$, en el k-ésimo tiempo de supervivencia no censurado, $k = 1, 2, \ldots, r-1$, donde n_j es el número de individuos en riesgo al tiempo t_j. Esta suma es simplemente $\sum_{j=1}^{k} 1/(n-j+1)$, que es el valor esperado de la k-ésima estadística de orden en una muestra de tamaño n de una distribución exponencial con parámetro uno.

Una debilidad de las gráficas basadas en los residuos es que no existen maneras objetivas de asegurar si hay alguna deficiencia en el modelo ajustado. En lugar de eso se utiliza un juicio informal para determinar, por ejemplo, si los puntos en una gráfica de riesgo log-acumulativo se desvían de la línea recta o si existe algún patrón en las gráficas de residuos contra los tiempos de supervivencia. Más aún, siempre hay incertidumbre acerca de cómo deberían verse las gráficas basadas en residuos si se está ajustando un modelo apropiado. Por ejemplo, aunque los residuos de martingala generalmente se prefieren sobre los residuos de Cox-Snell, se espera que las gráficas de los residuos de martingala contra el tiempo de supervivencia tengan un patrón, aún cuando se ajustó el modelo apropiado. Esto fue mostrado por Henderson y Milner, que propusieron que, sobre la gráfica de los residuos, se sobrepusieran los estimadores de la media esperada de los residuos para cada tiempo de supervivencia. De esta manera, los patrones que estén cercanos a la línea sobrepuesta no se interpretarán como deficiencias en el modelo.

En resumen, se recomienda utilizar en general los residuos de martingala y de devianza. Las gráficas que se recomienda utilizar son las de residuos contra las variables explicativas, contra los tiempos de supervivencia y las de índices. En un análisis más profundo, se deben calcular los residuos de Schoenfeld para cada variable explicativa en el modelo y graficarlos contra los tiempos de supervivencia. Sin embargo, nótese que las gráficas con patrones sistemáticos son las de residuos de Schoenfeld asociados a variables indicadoras utilizadas como factores en el modelo de regresión de Cox.

1.6.4 Identificación de observaciones influyentes.

Cuando se está valorando si un modelo es adecuado es importante determinar si una observación en particular tiene un impacto excesivo en las inferencias realizadas en base al modelo ajustado a un conjunto de datos de supervivencia observado. A las observaciones que tienen efecto en las inferencias basadas en el modelo se les llama *influyentes*.

Como un ejemplo, considérese un estudio de supervivencia en el cual se compara un tratamiento nuevo con el estándar. En tal comparación, será importante determinar si el riesgo de muerte con el tratamiento nuevo, relativo al estándar, se vio sustancialmente afectado por algún individuo. En particular, puede suceder que cuando no se toma en cuenta para el ajuste a un individuo, el riesgo relativo aumenta o disminuye de manera considerable. Si esto sucede, se necesitarán revisar cuidadosamente los registros de ese individuo.

Las conclusiones de un análisis de supervivencia frecuentemente se basan en los estimadores de los parámetros β en el modelo de regresión de Cox ajustado. Es, entonces, de particular interés examinar la influencia que tiene cada observación en estos estimadores. Esto se puede hacer examinando hasta qué punto los parámetros estimados en el modelo ajustado son afectados por la omisión de los datos relacionados a cada uno de los individuos dentro del estudio. En algunas circunstancias, los estimadores de cierto subconjunto de parámetros serán de especial importancia, como los parámetros asociados con

efectos de tratamientos. Entonces, el estudio de la influencia se limitará solamente a esos parámetros. En muchas ocasiones, será de interés la influencia que tiene cada observación en la función de riesgo estimada y será importante identificar las observaciones que influencian al conjunto completo de estimadores de los parámetros bajo el modelo.

En contraste con los modelos que se encuentran en el análisis de otro tipo de datos, como por ejemplo el modelo lineal general, no es fácil estudiar el efecto que produce el remover una observación de un conjunto de datos de supervivencia. Esto se debe principalmente a que la función de log-verosimilitud para el modelo de regresión de Cox no puede expresarse como la suma de cierto número de términos, en la cual cada término es la contribución de cada observación a la log-verosimilitud. En lugar de eso, la remoción de una observación afecta los conjuntos de riesgo sobre los cuales se suman las cantidades de la forma $e^{\beta' x}$. Esto significa que el diagnóstico de influencia es difícil de derivar así que a continuación se darán algunos resultados relevantes.

1.6.4.1 Influencia de las observaciones en uno de los parámetros.

Supóngase que se desea determinar si una observación en particular tiene un efecto inusual en $\hat{\beta}_j$, el estimador del j-ésimo parámetro, $j = 1, 2, \ldots, p$, en un modelo de regresión de Cox ajustado. Una manera de hacer esto sería ajustar el modelo a las n observaciones en el conjunto de datos y después ajustar el mismo modelo a los conjuntos de $n-1$ observaciones que se obtienen omitiendo cada una de las n observaciones. Se podrá determinar, entonces, cuál es el efecto que tiene en el parámetro el omitir cada una de las observaciones. Este procedimiento es computacionalmente caro, a menos que el número de observaciones sea muy pequeño, por lo que se utiliza una aproximación a la cantidad en la cual $\hat{\beta}_j$ cambia cuando se omite la i-ésima observación, para $i = 1, 2, \ldots, n$. Sea $\hat{\beta}_{j(i)}$ el valor del estimador del j-ésimo parámetro cuando se omite la i-ésima observación. Se han propuesto numerosas aproximaciones a la

cantidad $\hat{\beta}_j - \hat{\beta}_{j(i)}$, pero la considerada a continuación fue presentada por Cain y Lange en 1984.

Como ya es costumbre, t_i denotará el tiempo de supervivencia para el i-ésimo de los n individuos, y δ_i será un indicador de censura, que toma el valor cero si dicho tiempo de supervivencia está censurado y uno en otro caso. $\beta' \underline{x}_i$ será el predictor lineal para el i-ésimo individuo y defínase

$$a_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp\left(\hat{\underline{\beta}} \underline{x}_l\right)}{\sum_{l \in R(t_i)} \exp\left(\hat{\underline{\beta}} \underline{x}_l\right)}$$

donde x_{jl} es el valor de la j-ésima variable explicativa para el l-ésimo individuo en $R_{(t_i)}$, el conjunto de riesgo en t_i. Esta cantidad se utilizó en la definición del residuo de Schoenfeld y se presentó previamente en la ecuación (1.32).

Ahora, sea $\underline{d}_i$, $i = 1, 2, \ldots, n$, un vector de $p \times 1$ cuyo j-ésimo componente es

$$d_{ij} = \delta_i \left(x_{ji} - a_{ji}\right) + \exp\left(\hat{\underline{\beta}} \underline{x}_l\right) \sum_{t_j < t_i} \delta_j \frac{\left(a_{ji} - x_{ji}\right)}{\sum_{l \in R(t_i)} \exp\left(\hat{\underline{\beta}} \underline{x}_l\right)} \qquad (1.33)$$

para $j = 1, 2, \ldots, p$. Hay, por consiguiente, un vector $\underline{d}_i$ correspondiente a cada una de las n observaciones en el conjunto de datos. Nótese que el primer término en la expresión para d_{ij} es el residuo de Schoenfeld de la ecuación (1.31) asociado con la j-ésima variable explicativa en el modelo ajustado. Puede mostrarse que una aproximación a $\hat{\beta}_j - \hat{\beta}_{j(i)}$, el cambio en $\hat{\beta}_j$ al omitir la i-ésima observación, es el (i, j)-ésimo elemento de la matriz de $n \times p$

$$\Delta' V\left(\hat{\underline{\beta}}\right) \qquad (1.34)$$

donde $V(\underline{\hat{\beta}})$ es la matriz de varianza-covarianza de $p \times p$ del vector de estimadores de los parámetros en el modelo de regresión de Cox ajustado, y Δ' es la matriz de $n \times p$ cuya i-ésima fila consiste de los componentes del vector $\underline{d}_i$. Esta cantidad, a la que se le llama *delta-beta*, puede ser denotada por $\Delta_i\hat{\beta}_j \approx \hat{\beta}_j - \hat{\beta}_{j(i)}$. La expresión en la ecuación (1.34) es un poco complicada, pero tiene el mérito de que puede calcularse utilizando cantidades que están disponibles después de ajustar el modelo al conjunto completo de datos.

Las observaciones que tienen influencia sobre el estimador de algún parámetro en particular, por ejemplo el j-ésimo, serán aquellas cuyos valores de $\Delta_i\hat{\beta}_j$, las delta-betas para estas observaciones, son más grandes en valor absoluto que los de las otras observaciones en el conjunto de datos. Las gráficas de las delta-betas para cada variable explicativa en el modelo revelarán si hay observaciones cuya omisión cambiaría de manera significativa el valor del estimador del parámetro para esa variable en particular. En adición, una gráfica de los valores de $\Delta_i\hat{\beta}_j$ contra los tiempos de supervivencia proporciona información acerca de la relación entre el tiempo de supervivencia y la influencia.

Las delta-betas pueden estandarizarse dividiendo $\Delta_i\hat{\beta}_j$ por el error estándar de $\hat{\beta}_j$ para obtener una *delta-beta estandarizada*. La delta-beta estandarizada puede interpretarse como el cambio en el valor de la estadística de Wald, $\hat{\beta}/s.e.(\hat{\beta})$, al omitir la i-ésima observación. Como esta estadística puede usarse para determinar si un parámetro en particular tiene un valor significativamente diferente de cero, la delta-beta estandarizada puede utilizarse para informar cómo afecta a la significancia del parámetro el remover la i-ésima observación de la base de datos. Una vez más, una gráfica de índice es la forma más usual para observar las delta-betas estandarizadas.

La estadística en la ecuación (1.34) es una aproximación al verdadero cambio en el estimador del parámetro cuando la i-ésima observación se omite del

ajuste. La aproximación generalmente es adecuada en el sentido de que las observaciones que tienen influencia en el estimador de algún parámetro serán resaltadas. Sin embargo, se necesita estudiar el efecto real que tiene el omitir alguna observación en particular en las inferencias basadas en el modelo.

1.6.4.2 Influencia de las observaciones en los estimadores de un conjunto de parámetros.

Puede suceder que la estructura del modelo ajustado sea particularmente sensible a una o más observaciones del conjunto de datos. Tales observaciones pueden detectarse utilizando ciertos diagnósticos diseñados para resaltar las observaciones que influencian al conjunto completo de estimadores en el predictor lineal. Estos diagnósticos proveen información adicional a la información proporcionada por las delta-betas. En particular, el excluir cierta observación del conjunto de datos puede no tener una gran influencia en el estimador de un parámetro en particular y, por lo tanto, no será revelada por medio del estudio de las estadísticas delta-beta. Sin embargo, el cambio al remover tal observación en el conjunto de estimadores de los parámetros puede ser tal que la forma de la función de riesgo estimada o los valores de las estadísticas basadas en el modelo ajustado cambian marcadamente. Las estadísticas para valorar la influencia de las observaciones en el conjunto de estimadores de los parámetros también tienen la ventaja de que hay un sólo valor del diagnóstico para cada observación. Esto hace que sea más fácil utilizarlas que a los residuos de Schoenfeld o a las delta-betas.

Se han propuesto cierto número de diagnósticos para evaluar la influencia de cada observación en el conjunto de estimadores de los parámetros, a continuación se describirán dos de ellas.

Una manera de valorar la influencia de cada observación en el ajuste total de un modelo es examinar el cambio en el valor de menos dos veces el logaritmo de la máxima verosimilitud parcial, $-2\log\hat{L}$, bajo el modelo ajustado, al dejar

fuera alguna de las observaciones. Sea $-2\log L\left(\hat{\underline{\beta}}\right)$ el valor de la función de log verosimilitud maximizada cuando el modelo se ajusta a las n observaciones, y $-2\log L\left(\hat{\underline{\beta}}_{(i)}\right)$ el valor de la función de log-verosimilitud maximizada cuando se ajusta el modelo a las $n-1$ observaciones que quedan al omitir la i-ésima. El diagnóstico

$$2\left\{\log L\left(\hat{\underline{\beta}}\right)-\log L\left(\hat{\underline{\beta}}_{(i)}\right)\right\}$$

puede ser útil al estudiar la influencia.

Pettitt y Bin Daud (1989) mostraron que una aproximación a esta cantidad es

$$LD_i = \underline{d}_i' V\left(\hat{\underline{\beta}}\right)\underline{d}_i,$$

donde $\underline{d}_i$ es el vector de $p\times 1$ cuyo j-ésimo componente está dado en la ecuación (1.33) y $V\left(\hat{\underline{\beta}}\right)$ es la matriz de varianzas-covarianzas de $\hat{\underline{\beta}}$, el vector de estimadores de los parámetros. Los valores de esta estadística pueden obtenerse directamente de los términos utilizados para calcular las delta-betas para cada variable explicativa en el modelo. Una gráfica de índices de estas cantidades o contra los tiempos de supervivencia proporciona un resumen visual de los valores del diagnóstico. Las observaciones que tienen valores relativamente grandes de este diagnóstico son influyentes. No se recomienda realizar las gráficas contra las variables explicativas ya que suelen tener patrones determinados, aún si el modelo está correctamente ajustado.

Como en el método anterior cabe la duda de qué tan grande es grande, otro diagnóstico que puede utilizarse para valorar el impacto de cada observación en el conjunto de estimadores de los parámetros se basa en la matriz de $n\times n$

$$B = \Delta' V\left(\hat{\underline{\beta}}\right)\Delta,$$

donde Δ' es la matriz de $n \times p$ formada por los vectores $\underline{d}_i$, utilizada en la ecuación (1.34). Un argumento tomado de álgebra lineal muestra que los valores absolutos de los elementos del eigenvector estandarizado asociado con el mayor eigenvalor de la matriz B, es una medida de la sensibilidad del ajuste del modelo a cada una de las n observaciones en el conjunto de datos. Denotando este eigenvector por l_{max}, su i-ésimo elemento es una medida de la influencia de la i-ésima observación en el conjunto de los estimadores de los parámetros. El signo de esta cantidad es indiferente, así que las gráficas basadas en los valores absolutos, $|l_{max}|$ se recomiendan para uso general. Las gráficas de índices de estos valores, contra los tiempos de supervivencia y contra las variables explicativas en el modelo, pueden ser útiles en la valoración de la influencia.

Como l_{max} está estandarizado, los cuadrados de sus elementos suman uno. Las observaciones para las cuales los cuadrados de los elementos del eigenvector aportan una proporción sustancial a la suma total de los cuadrados serán las más influyentes. Los elementos más grandes de este vector corresponderán, entonces, a las observaciones que tienen mayor efecto en el valor de la función de verosimilitud. Un punto final es notar que, a diferencia de lo que sucede con otros diagnósticos, la gráfica de los elementos de l_{max} contra las variables explicativas no tienen un patrón determinado si el modelo ajustado es correcto. Esto significa que las gráficas de los valores absolutos de los elementos de l_{max} contra las variables explicativas serán útiles para evaluar si hay rangos de valores de las variables sobre las cuales el modelo no ajusta bien.

1.6.4.3 Tratamiento de las observaciones influyentes.

Una vez que se ha encontrado que ciertas observaciones son influyentes, es difícil decir qué se debe hacer con ellas. Esto depende en gran parte del contexto científico del estudio.

Cuando sea posible, debe revisarse el origen de las observaciones influyentes. Pueden existir errores al transcribir y al registrar datos categóricos y

numéricos. Si se encuentran los errores, se deberán corregir los datos y volver a repetir el análisis. Si los valores del tiempo de supervivencia o de las otras variables explicativas son imposibles y no se puede realizar la corrección, se deberá omitir tal dato de la base antes de repetir el análisis.

En muchas situaciones no será posible confirmar si los datos correspondientes a una observación influyente son válidos. Ciertamente, no se deben rechazar las observaciones rotundamente. Una alternativa es tener el análisis con y sin la observación influyente y dependiendo del problema determinar cuál se utiliza.

1.6.5 Probando el supuesto de riesgos proporcionales

Una suposición crucial al utilizar el modelo de Cox es la de riesgos proporcionales. Si los riesgos no son proporcionales significa que el componente lineal del modelo varía con el tiempo. Se debe, entonces, considerar cómo se puede examinar esta suposición críticamente.

Supóngase que los datos de supervivencia están agrupados de acuerdo a los niveles de uno o más factores y se encuentra el estimador de Kaplan-Meier de la función de supervivencia de los datos en cada grupo. Entonces, una gráfica de riesgos log-acumulada mostrará curvas paralelas si los riesgos son proporcionales a través de los diferentes grupos. Este método es informativo y simple de realizar si se tiene un número pequeño de grupos; sin embargo, si se tienen conjuntos de datos más estructurados se necesitará utilizar otro método.

La dependencia del tiempo del componente lineal de un modelo de riesgos proporcionales ajustado puede ser una consecuencia de la presencia en el modelo de variables explicativas que varían con el tiempo. En tales casos, se necesita un método que detecte alguna dependencia del tiempo de ciertas covariables, después de permitir el efecto de variables explicativas que se conoce, o se espera, sean independientes del tiempo. Esto sugiere utilizar un modelo que examine la validez de la suposición de riesgos proporcionales.

1.6.5.1 Prueba de riesgos proporcionales para el modelo de Cox

Los residuos de Schoenfeld pueden utilizarse para probar la suposición de riesgos proporcionales. Si esta suposición se cumple, la gráfica de estos residuos será una caminata aleatoria. Al contrario, supóngase que alguna variable, como un tratamiento, tiene un efecto positivo grande al principio pero va disminuyendo. El tratamiento puede tener influencia sobre cuántos pacientes sobreviven hasta cierto tiempo *t*, pero una vez que se "curan" no tiene influencia en la supervivencia más allá de *t*. En este caso, los riesgos proporcionales no se mantienen y los modelos ajustados subestiman el verdadero efecto del tratamiento para tiempos *t* pequeños, y lo sobreestiman para tiempos grandes. Si el tratamiento tiene un efecto benéfico, los residuos de Schoenfeld tendrán una tendencia negativa al principio seguida por una tendencia positiva tardía. Harrell sugiere utilizar la correlación entre los tiempos de supervivencia ordenados con este residuo como una prueba de riesgos no proporcionales.

Grambsch y Therneau mostraron después que el residuo de Schoenfeld reescalado puede corregir la correlación entre las covariables y ser más interpretable. Consideraron coeficientes que varían en el tiempo $\underline{\beta}(t) = \underline{\beta} + \underline{\theta} g(t)$, donde $g(t)$ es una función suave, usualmente el estimador Kaplan-Meier, $KM(t)$. Dada $g(t)$, ellos desarrollaron una prueba para $H_0 : \underline{\theta} = \underline{0}$ basada en un estimador de mínimos cuadrados generalizados para $\underline{\theta}$. Definiendo los residuos de Schoenfeld escalados como el producto de la inversa de la matriz de varianzas-covarianzas estimada para estos residuos. También desarrollaron un método gráfico. Mostraron por simulaciones Monte Carlo que un diagrama de dispersión suavizado de $\underline{\hat{\beta}}(t_k)$, el *k*-ésimo residuo de Schoenfeld reescalado más $\underline{\hat{\beta}}$, contra t_k revela la forma funcional de $\beta(t)$. Bajo H_0, se espera ver una función constante a través del tiempo.

Para realizar una prueba asintótica de ji-cuadrada con *p* grados de libertad para probar la hipótesis nula, se utiliza la siguiente estadística:

$$\left(\sum G_k r_k\right)' \left(\sum G_k V_k G_k\right)^{-1} \left(\sum G_k r_k\right)$$

donde r_k es la suma de tres elementos: el residuo de Schoenfeld del modelo verdadero, una variable aleatoria con media cero y la diferencia entre las medias ponderadas de las covariables bajo el modelo verdadero y el nulo; G_k es una matriz diagonal que depende de la función g utilizada y V_k la matriz de varianzas de los residuos de Shoenfeld.

En S-Plus la prueba de riesgos proporcionales arriba mencionada se realiza con la función cox.zph. La función g que utiliza es KM. La tabla obtenida es una matriz con una fila para cada variable y una última para la prueba global. Las columnas de la matriz contienen el coeficiente de correlación entre el tiempo de supervivencia transformado y los residuos Schoenfeld escalados (rho), la prueba ji-cuadrada (chisq) y el valor p de una prueba de dos colas. Para la prueba global no existe una correlación apropiada, por lo que en su lugar se coloca NA.

1.7 VARIABLES DEPENDIENTES DEL TIEMPO (TIME-DEPENDENT VARIABLES)

Hasta ahora se ha visto cómo modelar la dependencia de la función de riesgo en los valores de ciertas variables explicativas de un individuo. Cuando se incorporan variables explicativas en un modelo para datos de supervivencia, los valores que toman tales variables son los que se registraron al inicio del estudio y con ellos se evalúa el impacto que tienen sobre el riesgo de muerte.

En muchos estudios que generan datos de supervivencia se monitorea a los individuos mientras dura el estudio. Durante este periodo, se registran los valores de ciertas variables explicativas en intervalos regulares. Si se pueden tomar en cuenta los valores de estas variables conforme van evolucionando, puede ser posible proporcionar un mejor pronóstico del tiempo medio de vida del paciente.

Las variables cuyos valores cambian en el tiempo se conocen como *variables dependientes del tiempo.* A continuación se verá cómo se puede incorporar tales variables en un modelo utilizado para analizar datos de supervivencia. En este proceso, se utiliza el valor más reciente de la variable dependiente del tiempo en cada momento específico del procedimiento de modelación.

1.7.1 Tipos de variables dependientes del tiempo

Es útil considerar dos tipos de variables que cambian en el tiempo, variables *internas* y *variables externas.*

Las variables internas son aquellas que sólo se relacionan con un individuo en particular en el estudio y que sólo pueden medirse mientras el paciente está vivo. Como ejemplos pueden tomarse el conteo de glóbulos blancos en la sangre, presión arterial, etc. Por otro lado, las variables externas son aquellas variables dependientes del tiempo que no necesariamente requieren del paciente para existir.

Un tipo de variable externa es la que cambia de tal manera que se conoce qué valores tomará en el futuro. El ejemplo más obvio es la edad del paciente, ya que al saberse su edad en el tiempo origen se conocerá exactamente su edad en cualquier momento del futuro. Sin embargo hay otros ejemplos, como la dosis de una droga que varía de manera predeterminada durante el curso del estudio.

Otro tipo de variable externa es una que existe independientemente de cualquier individuo en particular, por ejemplo el nivel de dióxido de sulfuro en la atmósfera o la temperatura del aire. Los cambios en los valores de tales cantidades bien pueden tener un efecto en el tiempo de vida de los individuos, como pasa en estudios concernientes a pacientes con ciertos tipos de enfermedades respiratorias.

Estos diferentes tipos de variable dependiente del tiempo pueden introducirse dentro del modelo de riesgos proporcionales de Cox. El modelo resultante será llamado simplemente *modelo de regresión de Cox.*

1.7.2 Un modelo con variables dependientes del tiempo.

De acuerdo con el modelo de riesgos proporcionales de Cox descrito anteriormente, el riesgo de muerte al tiempo t para el i-ésimo de n individuos en un estudio puede escribirse como

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\} h_0(t)$$

donde x_{ji} es el valor base para la j-ésima variable explicativa, X_j, $j = 1, 2, \ldots, p$, para el i-ésimo individuo, $i = 1, 2, \ldots, n$, y $h_0(t)$ es la función de riesgo base. Generalizando este modelo a la situación en la cual algunas de las variables explicativas dependen del tiempo, se escribe $x_{ji}(t)$ para el valor de la j-ésima variable explicativa al tiempo t para el i-ésimo individuo. El modelo de regresión de Cox queda entonces

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}(t)\right\} h_0(t) \qquad (1.35)$$

En este modelo, la función de riesgo base $h_0(t)$ se interpreta como la función de riesgo para un individuo para el cual todas las variables son cero en el tiempo origen y se mantienen en este valor durante todo el tiempo.

Es importante notar que en el modelo dado en la ecuación (1.35) los valores de las variables $x_{ji}(t)$ dependen del tiempo t, así que el riesgo relativo $h_i(t)/h_0(t)$ también depende del tiempo. Esto significa que el riesgo de muerte al tiempo t no es proporcional al riesgo base y el modelo ya no es de riesgos proporcionales.

Para interpretar los parámetros β en este modelo, considérese el cociente de las funciones de riesgo al tiempo t para dos individuos, el r-ésimo y el s-ésimo. Esto está dado por

$$\frac{h_r(t)}{h_s(t)} = \exp\left\{\beta_1\left[x_{r1}(t) - x_{s1}(t)\right] + \beta_2\left[x_{r2}(t) - x_{s2}(t)\right] + \cdots + \beta_p\left[x_{rp}(t) - x_{sp}(t)\right]\right\}.$$

El coeficiente β_j, $j = 1, 2, \ldots, p$, puede interpretarse como el cociente de riesgos para dos individuos cuyos valores de la j-ésima variable explicativa al tiempo t difiere en uno, teniendo los dos individuos los mismos valores para todas las demás variables explicativas en ese tiempo.

La función de supervivencia para el i-ésimo individuo se obtiene, de la ecuación (1.4), de la función de riesgo integrada, y está dada por

$$S_i(t) = \exp\left\{\int_0^t \exp\left(\sum_{j=1}^{p} \beta_j x_{ji}(u)\right) h_0(u)\, du\right\}.$$

Esta función de supervivencia depende no sólo de la función de riesgo base $h_0(t)$, sino también de los valores de las variables dependientes del tiempo sobre el intervalo $(0,t)$. Más aún, no se mantiene el resultado de que $S_i(t)$ puede expresarse como una potencia de la función de supervivencia, $S_0(t)$. Esto significa que la función de supervivencia generalmente es difícil de obtener para cualquier individuo.

1.7.2.1 Ajustando el modelo de Cox

Cuando el modelo de regresión de Cox se extiende para incorporar variables dependientes del tiempo, la función de log-verosimilitud parcial, de la ecuación (3.5) puede generalizarse a

$$\sum_{i=1}^{n} \delta_i \left\{ \sum_{j=1}^{p} \beta_j x_{ji}(t_i) - \log \sum_{l \in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}(t_i) \right) \right\}, \qquad (1.36)$$

donde $R(t_i)$ es el conjunto de riesgo al tiempo t_i, el tiempo de muerte para el i-ésimo individuo en el estudio, $i = 1, 2, \ldots, n$, y δ_i es un indicador de censura que toma el valor cero cuando el tiempo de supervivencia del i-ésimo individuo está censurado y uno en otro caso. Esta expresión puede maximizarse para obtener los estimadores de los parámetros β.

Para utilizar la ecuación (1.35) en este proceso de maximización, los valores para cada una de las variables en el modelo deben conocerse en cada tiempo de muerte para todos los individuos en el conjunto de riesgo al tiempo t_i. Esto no es problema para las variables externas cuyos valores están predeterminados, pero pueden ser un problema para variables externas que existen independientemente de los individuos en el estudio y, ciertamente, para variables internas.

Para ilustrar este problema, considérese un ensayo de dos terapias de mantenimiento para pacientes que han sufrido un infarto al miocardio. El nivel de colesterol, X, para tales pacientes, será medido al momento en que el paciente es admitido en el estudio y, después, en intervalos regulares de tiempo. Es, entonces, más probable que el riesgo de muerte para el i-ésimo paciente al tiempo t, $h_i(t)$, esté influenciado por el valor de la variable explicativa X al tiempo t que al tiempo origen, $t = 0$.

Ahora supóngase que el i-ésimo individuo muere al tiempo t_i y que hay otros dos individuos, r y s, en el conjunto de riesgo al tiempo t_i. Supóngase también que el individuo r muere al tiempo t_r, donde $t_r > t_i$, y que el tiempo de supervivencia del individuo s, t_s, está censurado en algún momento después de t_r. La situación se ilustra gráficamente en la Figura 1.2. En esta figura, las líneas punteadas se refieren a los tiempos en que se mide la variable X.

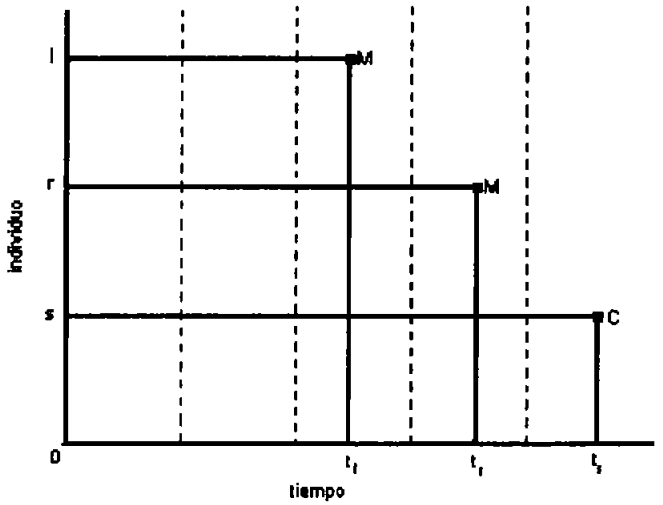

Figura 1.2 Tiempos de supervivencia de tres pacientes.

Si los individuos *r* y *s* son los únicos en el conjunto de riesgo al tiempo t_i, y *X* es la única variable explicativa medida, la contribución del *i*-ésimo individuo a la función de log-verosimilitud en la expresión (1.36) será

$$\beta x_i(t_i) - \log \sum_l \exp\{\beta x_l(t_i)\},$$

donde $x_i(t_i)$ es el valor de *X* para el *i*-ésimo individuo en su tiempo de muerte, t_i, y *l* en la suma toma los valores *i*, *r* y s. Esta expresión es igual a

$$\beta x_i(t_i) - \log\left\{e^{\beta x_i(t_i)} + e^{\beta x_r(t_i)} + e^{\beta x_s(t_i)}\right\}.$$

Esto muestra que se necesita el valor de la variable dependiente del tiempo, *X*, en el tiempo t_i para el *i*-ésimo individuo y para los individuos *r* y *s*. También se necesitará el valor de la variable *X* para los individuos *r* y *s* en t_r, el tiempo de muerte para el individuo *r*.

Para los términos en el modelo que son funciones explícitas del tiempo, tales como interacciones entre una variable o un factor y el tiempo, no hay dificultad al evaluar los valores de las variables dependientes del tiempo en cualquier

momento para cualquier individuo. Es más, usualmente es sencillo incorporar tales variables en el modelo de Cox cuando se utiliza software estadístico que tiene las facilidades para tratar con variables dependientes en el tiempo. Para otras variables, tales como el nivel de colesterol, se tienen que aproximar los valores de la variable dependiente en tiempos diferentes a los de las mediciones. Hay varias posibilidades.

Una opción es utilizar el último valor registrado de la variable antes del tiempo en el cual se necesita de su valor. Cuando para un individuo se han registrado los valores de la variable antes y después del momento en el que se requiere, se utiliza el valor del tiempo más cercano. Otra posibilidad es utilizar la interpolación entre valores consecutivos.

Claramente, la interpolación lineal no es opción cuando la variable dependiente del tiempo es categórica. Además, algunas variables categóricas pueden ser tales que sólo progresan a través de los niveles de la variable en una dirección en particular. Por ejemplo, cuando el estado de un paciente sólo puede cambiar de "bueno" a "intermedio" y de "intermedio" a "pobre". Como otro ejemplo, después de una biopsia, una variable asociada con la ocurrencia de un tumor tomará dos valores correspondientes a presencia o ausencia. Es muy poco probable que el estado cambie de "presente" a "ausente" en biopsias consecutivas.

2 APLICACIONES

2.1 Marco teórico.

La preeclampsia es un síndrome de origen desconocido que, aunque se desarrolla desde el inicio del embarazo, se manifiesta en el tercer trimestre de éste, y se caracteriza por hipertensión, alto contenido de proteína en la orina (proteinuria) y edema. Ocurre de modo más frecuente en el primer embarazo. La preeclampsia se puede complicar con condiciones que amenazan la vida de la embarazada e incluyen el desarrollo de convulsiones, insuficiencia hepática y falla renal. Es una de las causas principales de morbimortalidad materna en todo el mundo y también se le relaciona con resultados perinatales pobres que afectan al producto y que obedecen a una tasa alta de prematuros y a retraso de crecimiento intrauterino, pudiendo ocasionar bajo peso al nacer. Es por lo anterior, que cualquier esfuerzo para disminuir la preeclampsia o sus complicaciones asociadas, podría repercutir en el ámbito de la salud materno infantil.

La arginina, que es un aminoácido esencial, es indispensable para el desarrollo de la glándula mamaria y la lactación, en modelos animales experimentales. Por esta razón, es teóricamente posible que la deficiencia de arginina durante el embarazo pueda resultar en lactación inadecuada, que se puede remediar de manera muy sencilla, administrando arginina como suplemento.

Es muy común escuchar a los clínicos decir que la preeclampsia es una patología asociada a estados de malnutrición y es común también observar indicaciones de suplementación de la dieta de la embarazada, en especial en relación a las concentraciones disminuidas de proteína circulante. A pesar de que existen pocas evidencias de que soporten prevención del desarrollo de preeclampsia mediante manipulaciones de la dieta, algunos autores han enfatizado la utilidad del ajuste de la dieta para aliviar algunas de las manifestaciones secundarias de la enfermedad. Hasta el momento, no existe ningún estudio clínico controlado, con poder estadístico suficiente, que examine

el impacto de la suplementación de L-arginina en la incidencia de preeclampsia o sus complicaciones.

Existe suficiente información en modelos animales y en humanos que soporta el concepto de que la L-arginina[1] podría tener un efecto benéfico en la función renal de mujeres en riesgo de desarrollar preeclampsia o en aquellas que ya tienen instalada la enfermedad. La infusión de L-arginina reduce la presión arterial y reduce la resistencia vascular renal, ya que forma parte del ciclo del óxido nítrico que es un potente vasodilatador. La administración oral de L-arginina produce disminución en la presión arterial en sujetos con hipertensión leve de reciente diagnóstico en la primera semana de tratamiento.

Algunos investigadores han demostrado que la infusión intravenosa de L-arginina reduce las contracciones uterinas que aparecen asociadas al parto pretérmino. También utilizaron infusiones de L-arginina en mujeres con preeclampsia y en mujeres normales, en ambos casos lograron reducir la presión arterial materna y el efecto fue mayor en aquellas mujeres afectadas con eclampsia. Otros investigadores infundieron L-arginina en mujeres embarazadas complicadas con retraso en el crecimiento intrauterino y lograron incrementar la resistencia vascular uterina.

Con el objetivo de evaluar la eficacia de la suplementación con L-arginina y vitaminas antioxidantes en la incidencia de preeclampsia en una población de alto riesgo (mujeres embarazadas cuya presión arterial sea mayor o igual a 140/90 después de la semana 20 de gestación, sin proteinuria), el Instituto Nacional de Perinatología (INPer) realizó un estudio doble ciego, que consistió en dar un suplemento con forma de barras a cada mujer, dividiendo así a la población de manera aleatoria en 3 grupos, de acuerdo al tipo de barra alimenticia que se les administra. Las mujeres del Grupo 1 consumían barras placebo; las del Grupo 2, barras con vitaminas y L-arginina y las del Grupo 3, barras con sólo vitaminas. Cada dos semanas las pacientes deberían acudir a una visita de control para entregarles las barras, hacerles análisis, pesarlas y controlar su dieta.

---

[1] La L en L-arginina indica que el aminoácido ocupa una posición lateral dentro de la cadena de proteínas.

Una consecuencia de la preeclampsia es el término prematuro del embarazo, lo cual afecta también la salud del feto. Existen varios objetivos en este estudio, una en particular es ver si el término del embarazo es diferente en cada grupo. Por medio de modelos de análisis de sobrevivencia se analiza si en alguno de los tres el tiempo de gestación es mayor y cuáles son las variables que de alguna manera están más relacionadas con el término temprano del embarazo.

2.2 Análisis

Se tiene una base de datos que contiene variables posiblemente relacionadas con la preeclampsia, como lo son: niveles de sustancias en orina, presión arterial, desarrollo de la enfermedad y otras variables secundarias. Cada variable está registrada para cada una de las visitas que tuvo cada paciente. El número de visitas varía entre una y ocho, más la visita final (parto).

Se aceptó dentro del estudio a las pacientes embarazadas primigestas con 20 semanas o más de gestación y con producto único. Todas las mujeres enroladas debían continuar y terminar su embarazo en el INPer. La edad gestacional máxima para admitir a una mujer en el estudio fue de 34 semanas. Debían tener presión arterial sistólica mayor o igual a 140 mmHg y menor a 160 mmHg, y presión diastólica mayor o igual a 90 mmHg y menor a 110 mmHg, conforme a ciertos métodos establecidos. No debían presentar proteinuria. Los criterios de exclusión fueron los siguientes: Gestación múltiple, malformación fetal mayor, hipertensión preexistente, diabetes mellitus, enfermedad autoinmune, cáncer de cualquier tipo o historia de cáncer en parientes en primer grado y enfermedad materna preexistente que requirió medicación.

De acuerdo con ciertos cálculos realizados, el Instituto requería un total de 528 pacientes para distribuir en 3 grupos de 176 sujetos cada uno. Sin embargo, al momento de comenzar a realizar la presente tesis, el INPer nos proporcionó una base de datos conformada por 55 mujeres, de las cuales 7 no tenían ningún dato registrado. En general, existen valores faltantes en varias de las visitas.

Para empezar, se realizó una gráfica de los estimadores KM de las funciones de sobrevivencia, sin censura, para 48 mujeres (Figura 2.1).

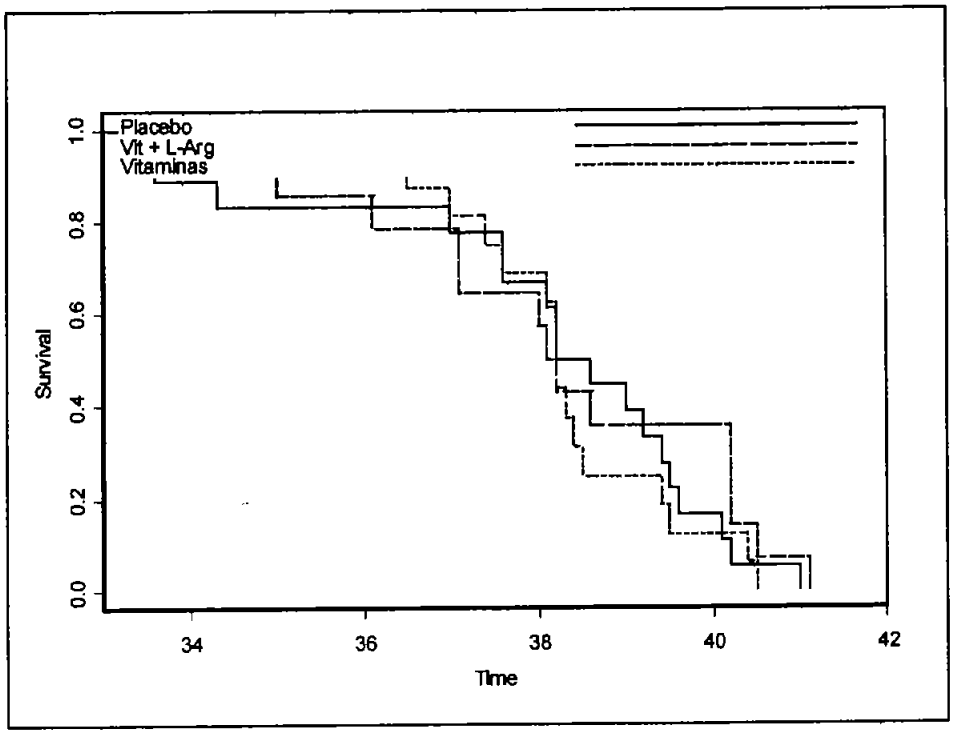

Figura 2.1. Estimadores KM de las funciones de sobrevivencia para los tres grupos.

Como puede observarse, no existe una diferencia marcada entre las curvas de sobrevivencia de los tres grupos.

Al realizar la prueba de log-rangos (log-rank test) se obtuvo un estadístico de prueba con valor 0.88, que, al ser comparado con el cuantil de una ji-cuadrada con 2 grados de libertad (5.99, α=0.05), nos indica que no existe evidencia suficiente para rechazar la hipótesis de que no hay diferencias entre los tres grupos.

A continuación, con una muestra inicial de 30 mujeres, que tenían registrada más de una visita, se realizó un análisis de sobrevivencia sin censura en S-Plus, ajustando modelos de riesgos proporcionales de Cox, y utilizando principalmente las variables relacionadas con la orina, sangre y con la presión. Todas las variables no categóricas se estandarizaron restando al valor de la

última visita registrada el de la primera y dividiendo después entre el valor de la primera.

Las variables consideradas fueron:

- *Generales:*
  - Grupo de tratamiento (*bar_id*). Variable categórica que toma los valores 1, 2 y 3, de acuerdo a la descripción realizada anteriormente.
  - Tiempo de sobrevivencia (*wog_mf*). Semana del parto.
  - Tiempo que duró la paciente dentro del estudio (*wog_dif*).
  - Edad de la paciente (*age*).
  - Diagnóstico final (*dx_mf*). Variable categórica que toma los valores N, si la paciente terminó normalmente su embarazo, o P, si la paciente presentó preeclampsia.
  - Censura (*censura*). Toma el valor 0 si se tiene un dato censurado y 1 si no, es decir, si el embarazo llega a término.

- *Relacionadas con la presión arterial:*
  - Presión sistólica media (*sbpm*).
  - Presión diastólica media (*dbpm*).

- *Relacionadas con la orina:*
  - Nitrógeno uréico sanguíneo (*bun*)
  - Glucosa (*glu*)
  - Creatinina (*crea*)
  - Ácido úrico (*ua*)
  - Ácido úrico categorizado (*uacat*). Toma el valor 1 si el nivel de ácido úrico (ua) es mayor que el percentil 75 y 0 si es menor o igual.

- *Relacionadas con la sangre:*
  - Lipoproteína de alta densidad (*hdl*)
  - Colesterol total (*tc*)
  - Triglicéridos (*tg*)
  - Proteína total (*tp*)

Primero se realizó el análisis utilizando las variables generales y las relacionadas con la presión, en S-Plus. Aunque por definición de preeclampsia (presión alta y proteinuria) las variables sbpm y dbpm deberían ser importantes, en los modelos de riesgos proporcionales considerados no resultaron significativas, por lo que los modelos obtenidos no se presentan aquí.

A continuación se ajustaron varios modelos de riesgos proporcionales, utilizando las variables generales y las relacionadas con la orina y con la sangre. Se inició con modelos que incluían solamente una de las variables de orina, sangre y bar_id y se fueron agregando variables e interacciones hasta obtener el siguiente modelo.

Modelo 1

```
      *** Cox Proportional Hazards ***
Call:
coxph(formula = Surv(wog.mf) ~ bar.id + ua + hdl + uacat:hdl + wog.dif +
     bar.id:dx.mf, data = orina, na.action = na.exclude, method = "efron", robust
     = F)

n=30
```

	coef	se(coef)	z	p	exp(coef)	lower .95	upper .95
bar.id1	-0.1740	0.4126	-0.4210	0.6700	0.8404	0.374	1.887
bar.id2	0.2480	0.2839	0.8730	0.3800	1.2812	0.734	2.235
ua	5.6390	2.0308	2.7770	0.0055	281.1600	5.252	15051.111
hdl	2.2130	0.8164	2.7110	0.0067	9.1454	1.846	45.307
wog.dif	-0.1210	0.0467	-2.5990	0.0093	0.8857	0.808	0.971
uacat:hdl	-2.8630	1.0020	-2.8570	0.0043	0.0571	0.008	0.407
bar.id1dx.mf	1.2230	0.6803	1.7970	0.0720	3.3960	0.895	12.884
bar.id2dx.mf	0.6960	0.7352	0.9460	0.3400	2.0053	0.475	8.473
bar.id3dx.mf	0.9380	0.6696	1.4010	0.1600	2.5547	0.688	9.491

```
Rsquare= 0.676   (max possible= 0.993 )
Likelihood ratio test= 33.8  on 9 df,   p=0.0000967
Wald test            = 21    on 9 df,   p=0.0127
Score (logrank) test = 34    on 9 df,   p=0.0000885
```

Dados los valores de los estadísticos de prueba y la significancia de las variables, parece ser un buen modelo para ajustar los datos, sin embargo, al realizar las gráficas de residuos de devianza se identificó una observación con un

residuo muy alto en la variable hdl (Figura 2.2). Tal observación es la paciente con el registro número 106.

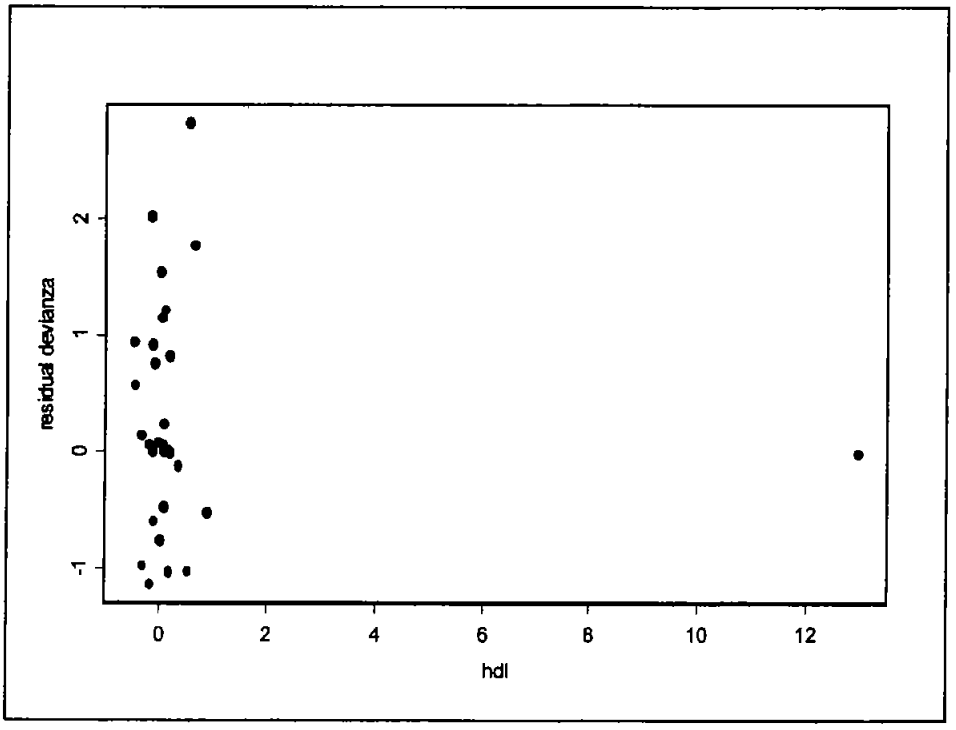

Figura 2.2. Residuos de devianza vs. hdl para el Modelo 1.

Al analizar la base de datos se encontró que el valor de la variable hdl para la primera visita de la paciente 106 (12.98) está fuera del rango de valores que, sin esta observación, va de -0.4477 a 0.9206, por lo que se pensó que el valor podría estar equivocado, tal vez un error de dedo. Para evitar suponer el valor real y a la vez para comprobar si la observación es influyente, se eliminó de la base de datos y se volvió a correr el Modelo 1, obteniéndose los siguientes resultados (Modelo 2).

Modelo 2 (Modelo 1 sin observación 106)

```
      *** Cox Proportional Hazards ***
Call:
coxph(formula = Surv(wog.mf) ~ bar.id + ua + hdl + uacat:hdl + wog.dif +
     bar.id:dx.mf, data = orina, na.action = na.exclude, method = "efron", robust
     = F)

n=29
```

	coef	se(coef)	z	p	exp(coef)	lower .95	upper .95
bar.id1	0.0705	0.4524	0.1560	0.8800	1.0700	0.442	2.600
bar.id2	0.3522	0.2998	1.1750	0.2400	1.4200	0.790	2.560
ua	7.0032	2.2703	3.0850	0.0020	1100.0000	12.900	94169.420
hdl	2.3437	0.8132	2.8820	0.0039	10.4000	2.120	51.290
wog.dif	-0.1389	0.0498	-2.7910	0.0053	0.8700	0.789	0.960
uacat:hdl	-9.8316	5.4183	-1.8150	0.0700	0.0001	0.000	2.200
bar.id1dx.mf	1.0360	0.6882	1.5050	0.1300	2.8200	0.731	10.860
bar.id2dx.mf	1.2466	0.8666	1.4380	0.1500	3.4800	0.636	19.010
bar.id3dx.mf	1.2892	0.7257	1.7760	0.0760	3.6300	0.875	15.050

```
Rsquare= 0.632   (max possible= 0.993 )
Likelihood ratio test= 29    on 9 df,   p=0.000642
Wald test            = 21.2  on 9 df,   p=0.0117
Score (logrank) test = 31.7  on 9 df,   p=0.000227
```

Como puede notarse, los valores de los parámetros cambian con respecto al Modelo 1, pero no de manera significativa. Los coeficientes para las variables ua y hdl son significativamente diferentes de cero. Para la primera el valor del coeficiente es 7.0032, lo que significa que mientras más aumente el ácido úrico entre la primera y la última visita, más riesgo tiene la paciente de terminar prematuramente su embarazo. Los valores de la variable ua están en el intervalo (-0.1578, 0.5266), cuya longitud (rango) es de 0.7 aproximadamente; por lo que el incremento en el riesgo de parto prematuro, cuando la variable ua aumenta en 0.175 (la cuarta parte del rango), es de $\exp(7.0032 * 0.175) = 3.406073$. Es decir, el incremento en el riesgo al aumentar la variable ua en 0.175, quedando los valores de todas las demás variables iguales, es de 3.406073.

Como es difícil interpretar las interacciones entre dos variables continuas, se optó por integrar en el modelo la interacción entre uacat y hdl, ya que, al ser la primera categórica, facilita la interpretación. En este modelo la variable uacat tomó el valor 1 si el nivel del ácido úrico era mayor a 0.2955340 y 0 si era menor o igual. El valor del coeficiente de la interacción (-9.8316) indica que cuando la

diferencia entre los niveles del ácido úrico entre la primera y la segunda visita es grande (uacat=1), al aumentar la lipoproteína de alta densidad (hdl) el riesgo de tener un parto prematuro disminuye. El que el coeficiente sea negativo indica que para las mujeres con niveles altos de ácido úrico es mejor tener también altos de hdl. Esta interacción también es significativamente diferente de cero.

Cuando la variable uacat es igual a 0, los valores de hdl están en el intervalo (-0.4477,0.9206), cuya longitud es de 1.4 aproximadamente; por lo que el incremento en el riesgo de parto prematuro, cuando la variable hdl aumenta en 0.35 (la cuarta parte del rango), es de $\exp(2.3437 * 0.35) = 2.27117$.

El coeficiente de la variable wog.dif también es significativo y toma un valor de -0.1389, por lo tanto, mientras más tiempo dura la paciente dentro del estudio menor es su riesgo de parto prematuro. Esto es lógico, ya que al estar dentro del estudio su alimentación, presión, etc., estaban más controladas.

A diferencia de lo que sucede en el Modelo 1, la interacción marginalmente significativa se da entre el Grupo 3 y el diagnóstico, el valor del coeficiente (1.2892) indica que, dentro de las pacientes en el tercer grupo, las que presentaron preeclampsia tienen mayor riesgo de interrumpir su embarazo prematuramente.

Con el objetivo de revisar que el modelo ajustara correctamente a los datos se realizaron las siguientes gráficas de residuos, con las que se confirma el buen ajuste del modelo. Se omitieron las gráficas de los residuos de martingala, por ser muy parecidas a las de los residuos de devianza.

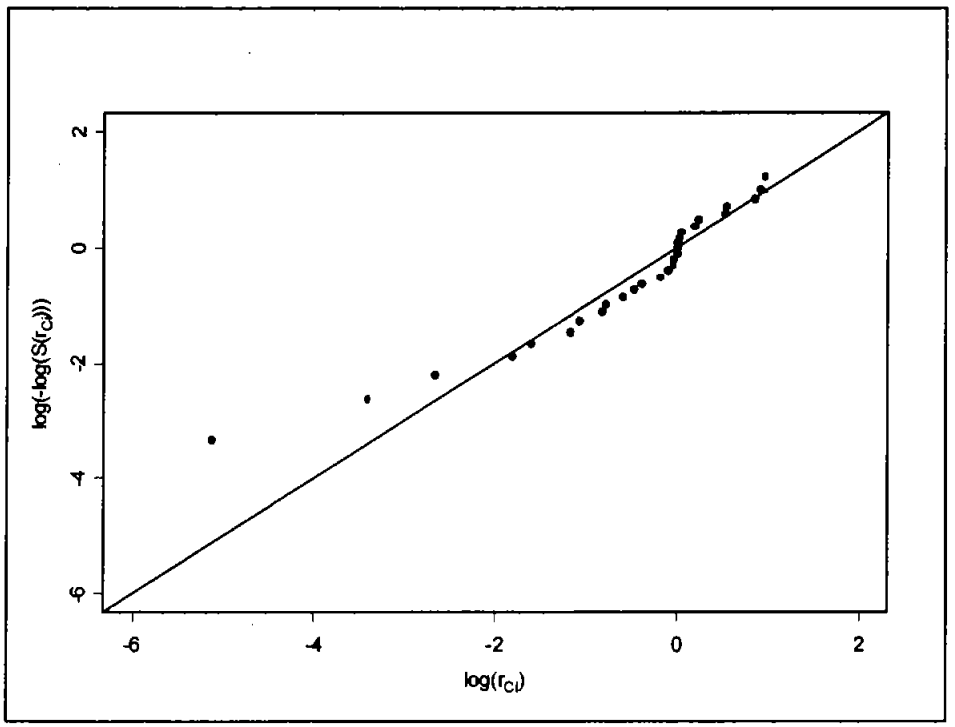

Figura 2.3. Gráfica del riesgo log-acumulativo para los residuos de Cox-Snell del Modelo 2.

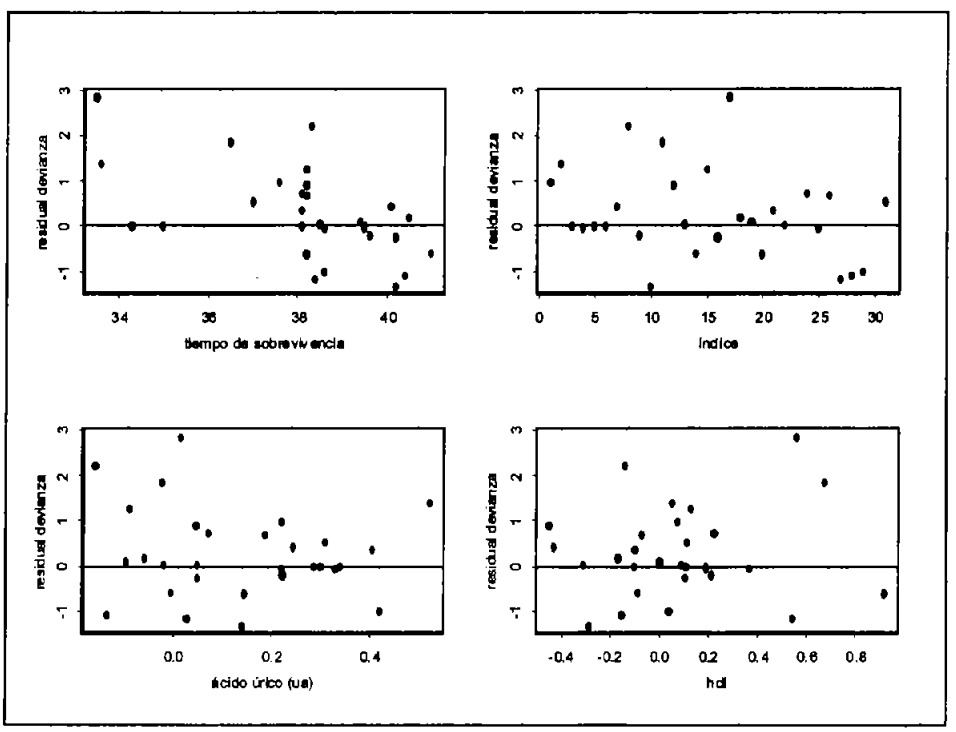

Figura 2.4. Residuos de devianza para el Modelo 2.

Observamos de la Figura 2.3 que, aunque los residuos no caen precisamente sobre la identidad, no están tan lejos de esta, así que el Modelo 2 ajusta razonablemente a los datos.

Los residuos de devianza deben de ser simétricos alrededor del cero si el modelo es adecuado. Como puede observarse en la Figura 2.4 hay tres valores extremos (para las 4 gráficas), que corresponden a las pacientes 73, 82 y 94, que presentaron diagnóstico normal, siendo las 2 primeras del Grupo 3 y la última del Grupo 1. Fuera de eso, parece que el modelo ajusta bien a los datos. Para revisar si los valores extremos son observaciones influyentes se obtuvieron las estadísticas delta-beta para cada una de las variables incluidas en el Modelo 2.

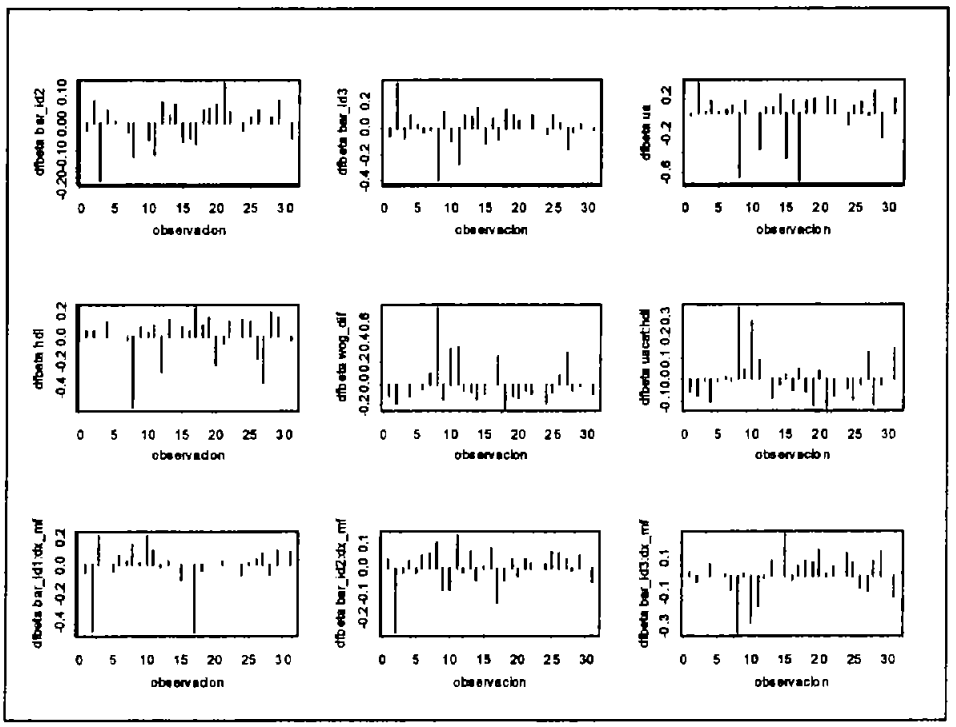

Figura 2.5. Delta-betas para detectar observaciones influyentes en los 9 estimadores de los coeficientes correspondientes a las variables explicativas para el Modelo 2.

La Figura 2.5 muestra que los cambios en los coeficientes de regresión son, cuando muy grandes, de ± 0.6 desviaciones estándar. Como la mayoría de las desviaciones estándar de los coeficientes son menores que uno, y las más grandes son de 2 y 5, se puede concluir que no hay observaciones influyentes.

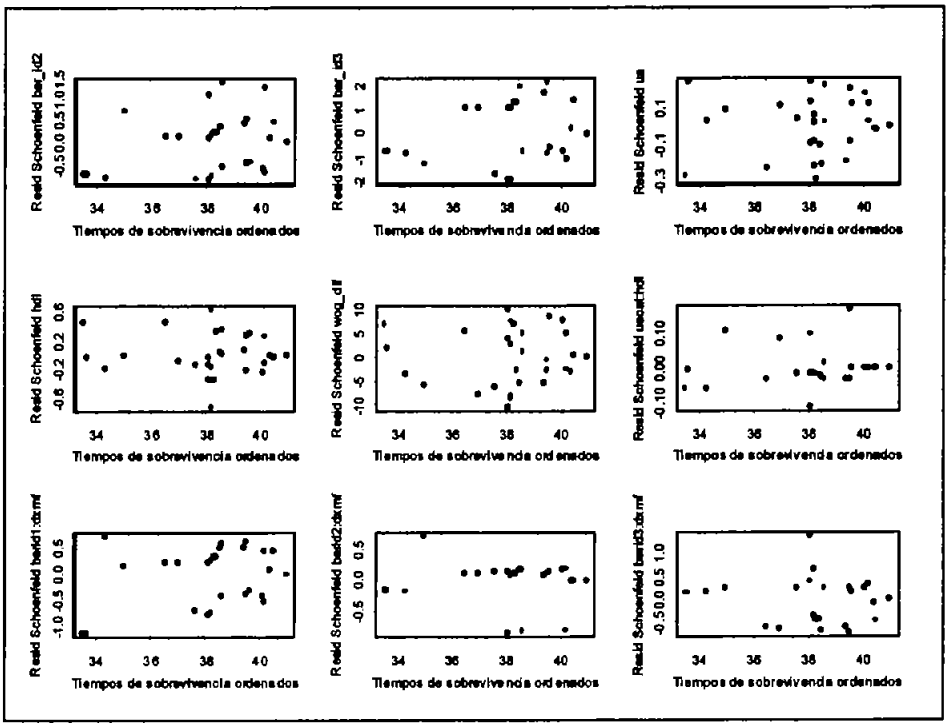

Figura 2.6. Residuos Schoenfeld contra los tiempos de sobrevivencia ordenados para el Modelo 2.

Para probar el supuesto de riesgos proporcionales se utilizó la gráfica de los residuos Schoenfeld contra los tiempos de sobrevivencia ordenados (Figura 2.6). Como puede observarse, hay cuatro puntos un poco separados de los demás, pero corresponden a los residuos correspondientes a las pacientes 12, 30, 62 y 94 que tienen 4 tiempos de sobrevivencia más pequeños. Por lo tanto, como no existe ningún patrón en las gráficas, parece que el supuesto de riesgos proporcionales es apropiado.

	rho	chisq	p
bar.id1	0.056	0.074	0.786
bar.id2	0.065	0.134	0.714
ua	0.154	1.496	0.221
hdl	-0.052	0.096	0.757
wog.dif	0.035	0.048	0.827
uacat:hdl	0.061	0.104	0.747
bar.id1dx.mf	0.116	0.481	0.488
bar.id2dx.mf	-0.036	0.034	0.855
bar.id3dx.mf	-0.105	0.344	0.558
GLOBAL	NA	4.880	0.845

Tabla 2.1. Prueba de Grambsch y Therneau para probar el supuesto de riesgos proporcionales en el Modelo 2.

Los resultados de la prueba de constancia de los coeficientes basada en los residuos Schoenfeld reescalados (Tabla 2.1) indica que el supuesto de riesgos proporcionales se satisface para todas las variables en el modelo, con todos los valores p mayores a 0.221. La Figura 2.7 comprueba el supuesto de riesgos proporcionales, ya que las líneas que suavizan todas las gráficas son casi planas.

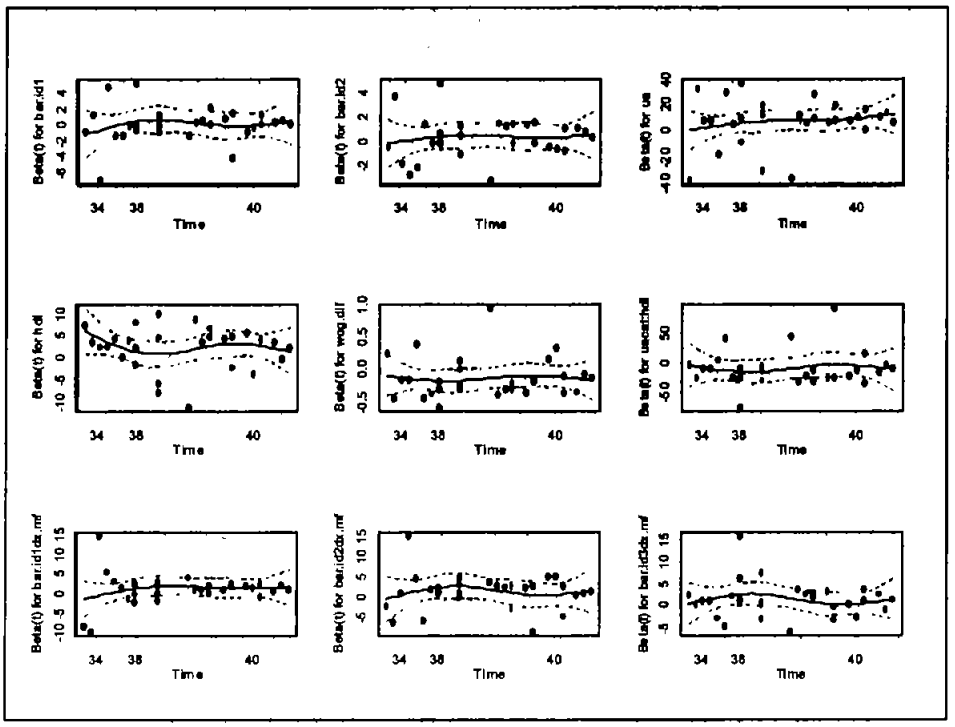

Figura 2.7. Gráficas de diagnóstico de la constancia de los coeficientes en el Modelo 2. Cada gráfica corresponde a un coeficiente del modelo contra el tiempo de sobrevivencia ordenado. Se muestra una línea de suavizamiento junto con las bandas de ± 2 desviaciones estándar.

En conclusión, puede considerarse al Modelo 2 como un modelo que ajusta adecuadamente a los datos.

Modelo 3

Hasta este momento se trabajó únicamente con datos sin censura, por lo que el número de individuos utilizados para ajustar el modelo anterior fue muy pequeño. A continuación se ajustará otro modelo de Cox utilizando datos con múltiples registros (multiple-record data), considerando censura, para hacer un análisis de variables dependientes del tiempo. El siguiente análisis y gráficas se realizaron en S-Plus.

Para realizar este tipo de análisis se tuvo que modificar un poco la base de datos, de manera que de cada paciente se obtuvieran varios registros, dependiendo del número de visitas que tuvo. A continuación se presentan varios ejemplos.

Supóngase que la paciente número 101 tiene registradas 2 variables, x_1 y x_2, para 2 visitas, la primera en la semana 36 y la segunda en la 40, donde terminó el embarazo (supóngase que el tiempo origen es la semana 20, ya que es la mínima semana en la que las mujeres pueden entrar al estudio), de la siguiente manera:

x_1 = 17 y x_2 = 22 durante (20,36],

x_1 = 12 y x_2 = 21 durante (36,40], como llegó a término el embarazo, censura = 1.

Esto queda registrado así:

file	wog_emp	wog_term	x_1	x_2	censura
101	20	36	17	22	0
101	36	40	12	21	1

También puede pasar lo siguiente:

x_1 = 17 y x_2 = 22 durante (20,34],

x_1 = 12 y x_2 = 21 durante (34,36], y no se tienen más visitas, pero se sabe que la última no fue el parto.

En este caso el registro queda así:

file	wog_emp	wog_term	x_1	x_2	censura
101	20	34	17	22	0
101	34	36	12	21	0

Como puede observarse, en lugar de tener un tamaño de muestra de tamaño 29, como fue en el Modelo 2, se tendrá una base mucho más grande. Además, en este caso, se tomó el valor no estandarizado de las variables.

Una vez modificada la base, se probaron varios modelos, utilizando las variables generales, de orina y sangre. Dentro de todos los modelos, el siguiente (Modelo 3) es el que resultó más significativo.

```
      *** Cox Proportional Hazards ***
Call:
coxph(formula = Surv(wog.emp, wog.term, censura, type = "counting") ~ bar.id +
     uacat + hdl + uacathdl + tc, data = basecnva, na.action = na.exclude, method
     = "efron", robust = F)

  n=124

                coef se(coef)       z       p exp(coef) lower .95 upper .95
    bar.id1  -0.2582   0.3911 -0.6600  0.5100    0.7720     0.359     1.660
    bar.id2  -0.1203   0.2633 -0.4570  0.6500    0.8870     0.529     1.490
      uacat   2.6282   1.2949  2.0300  0.0420   13.8490     1.094   175.240
        hdl   0.0768   0.0385  1.9980  0.0460    1.0800     1.001     1.160
   uacathdl  -0.0801   0.0429 -1.8680  0.0620    0.9230     0.849     1.000
         tc  -0.0200   0.0101 -1.9860  0.0470    0.9800     0.961     1.000

Rsquare= 0.056   (max possible= 0.407 )
Likelihood ratio test= 7.17  on 6 df,   p=0.306
Wald test            = 6.42  on 6 df,   p=0.378
Score (logrank) test = 7.37  on 6 df,   p=0.288
```

Aunque la variable bar_id (Grupo) no es significativa, los coeficientes nos indican que las mujeres que están dentro del grupo 1 (placebo) tienen mayor riesgo de terminar su embarazo prematuramente, ya que el valor del coeficiente es 0. El grupo que tiene el menor riesgo es el 2 (Vitaminas + L-arginina), ya que el valor del coeficiente es -0.2583, por lo que se le considera factor protector.

Si observamos el valor del coeficiente correspondiente a la interacción entre uacat y hdl, -0.0801, se nota que, cuando se tienen niveles altos tanto del ácido úrico como de hdl, el riesgo de parto prematuro disminuye. Ahora, si la variable hdl tomara el valor 0, se toma en cuenta el valor del coeficiente de la variable uacat, 2.6282, que resulta significativo y nos indica que cuando el nivel de ácido úrico es muy alto, mayor a 4.26 (el percentil 75), el riesgo de tener un parto prematuro aumenta. Cuando la variable uacat es igual a 1, se considera el coeficiente de hdl, 0.0768, que también es significativo e indica que al aumentar el valor de esta variable el riesgo de terminar prematuramente su embarazo no aumenta en gran medida.

En este modelo se incluyó a la variable tc (colesterol total), resultando esta significativa y un factor protector para el parto prematuro. Esto no resulta lógico, sin embargo, al no tener más datos no se puede comprobar si en realidad esto sucede.

A continuación se realizaron las gráficas de residuos para comprobar el buen ajuste de este modelo.

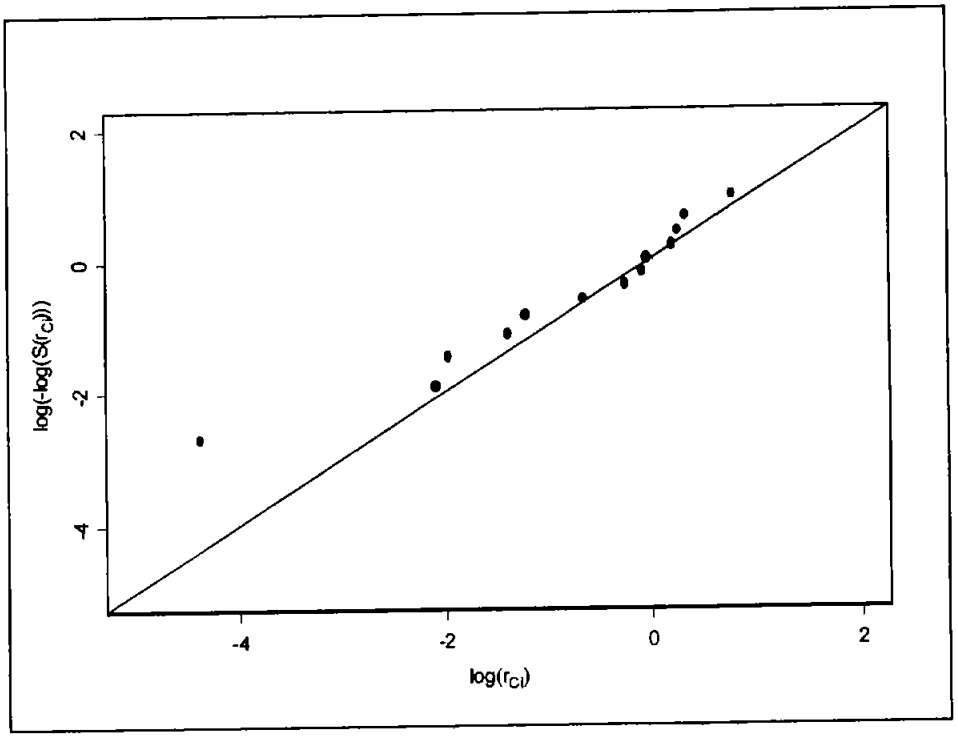

Figura 2.8. Gráfica del riesgo log-acumulativo para los residuos de Cox-Snell del Modelo 3.

Como puede observarse en la Figura 2.8, aunque los puntos no quedan exactamente sobre la recta, tampoco se separan mucho de ésta, por lo que puede considerarse que el modelo ajusta razonablemente a los datos. Cabe aclarar que en la gráfica sólo aparecen los puntos correspondientes a los registros no censurados, es por eso que sólo hay 16 puntos. Esto ocurre también en las sigueintes gráficas.

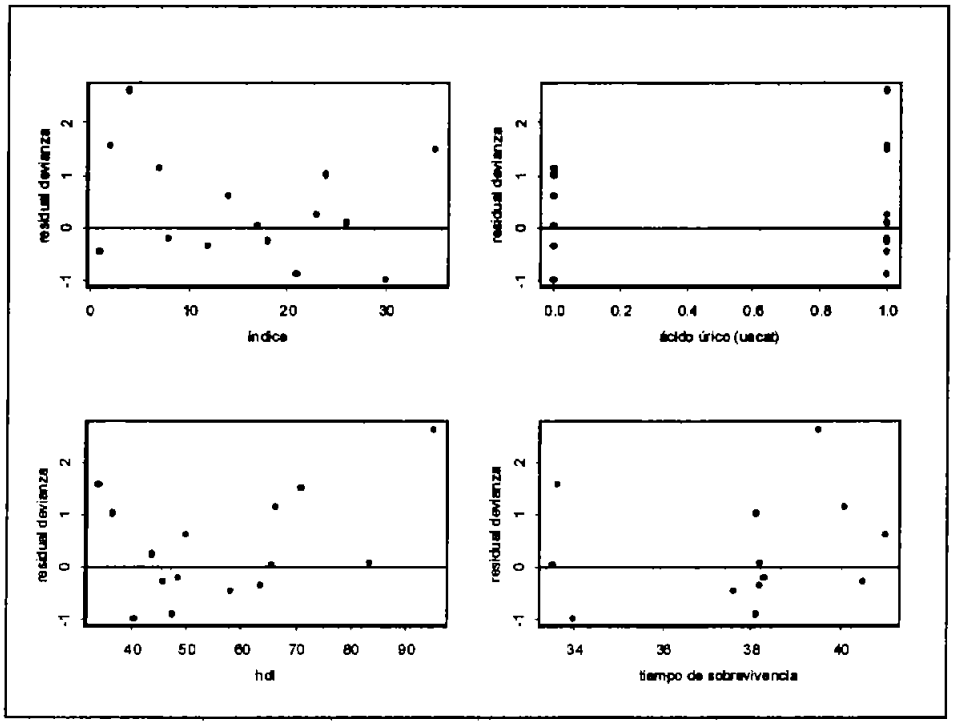

Figura 2.9. Residuos de devianza para el Modelo 3.

En las gráficas de los residuos de devianza (Figura 2.9) hay cuatro valores extremos, que son los residuos correspondientes a las pacientes números 12, 30, 66 y 124. Como los residuos no siguen ninguna tendencia en particular se confirma el buen ajuste del modelo.

A continuación se presentan las gráficas de los valores de las delta-betas para revisar si alguno de los valores extremos es una observación influyente.

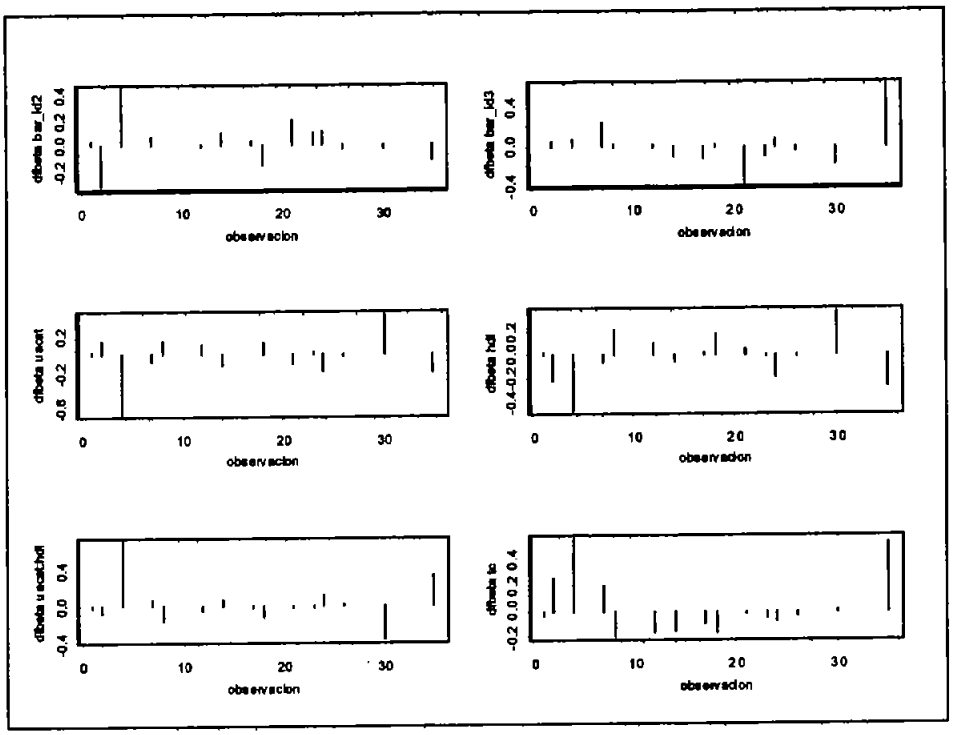

Figura 2.10. Delta-betas para detectar observaciones influyentes en los 9 estimadores de los coeficientes correspondientes a las variables explicativas para el Modelo 3.

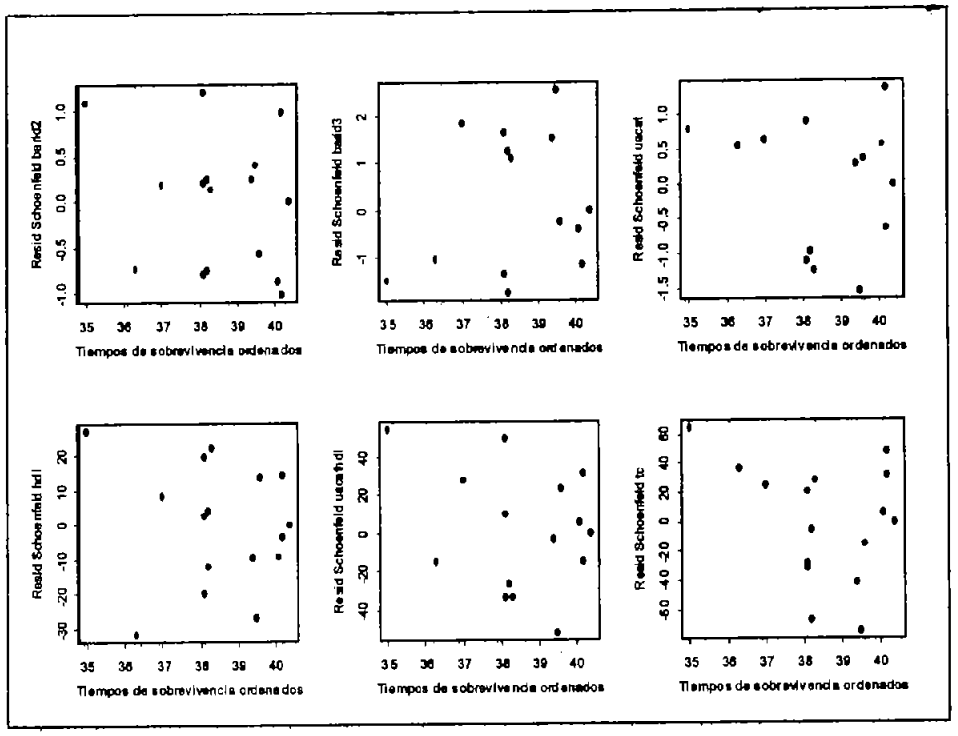

Figura 2.11.Residuos Schoenfeld contra los tiempos de sobrevivencia ordenados para el Modelo 3.

Al igual que en el Modelo 2 (Figura 2.10), ninguna de las observaciones tiene una delta-beta mayor a ± 0.6 desviaciones estándar de los coeficientes, y dado que el mayor valor de estas desviaciones es de 1.2, al eliminar alguna de estas observaciones los valores de los coeficientes no cambiarían mucho, por lo que no se les considera influyentes.

Como se realizó anteriormente para el Modelo 2, se utilizaron las gráficas de los residuos Schoenfeld (Figura 2.11) para probar el supuesto de riesgos proporcionales en el Modelo 3. Como puede observarse no existe ningún patrón en las gráficas, por lo que aparentemente el supuesto de riesgos proporcionales es apropiado.

	rho	chisq	p
bar.id1	-0.240	0.962	0.327
bar.id2	-0.151	0.499	0.480
uacat	0.489	4.399	0.036
hdl	0.475	3.580	0.059
uacathdl	-0.515	5.200	0.023
tc	-0.469	4.950	0.026
GLOBAL	NA	8.260	0.220

Tabla 2.2. Prueba de Grambsch y Therneau para probar el supuesto de riesgos proporcionales en el Modelo 3.

Al realizar la prueba de consistencia para los coeficientes (Tabla 2.2) el supuesto de riesgos proporcionales se rechaza para uacat, uacathdl y tc, aunque de manera global no se rechaza. Al observar las gráficas de la Figura 2.12, aparentemente ninguna de las líneas de suavizamiento está plana, sin embargo, en todas, excepto para tc, la línea con pendiente cero e intersección en el origen está incluida dentro de las bandas de confianza. Posiblemente, si se cambiara el nivel de suavizamiento de las líneas podrían verse más planas, pero no se pudo hacer esto debido a que S-Plus las grafica así por default y no hay opción para cambiarlas.

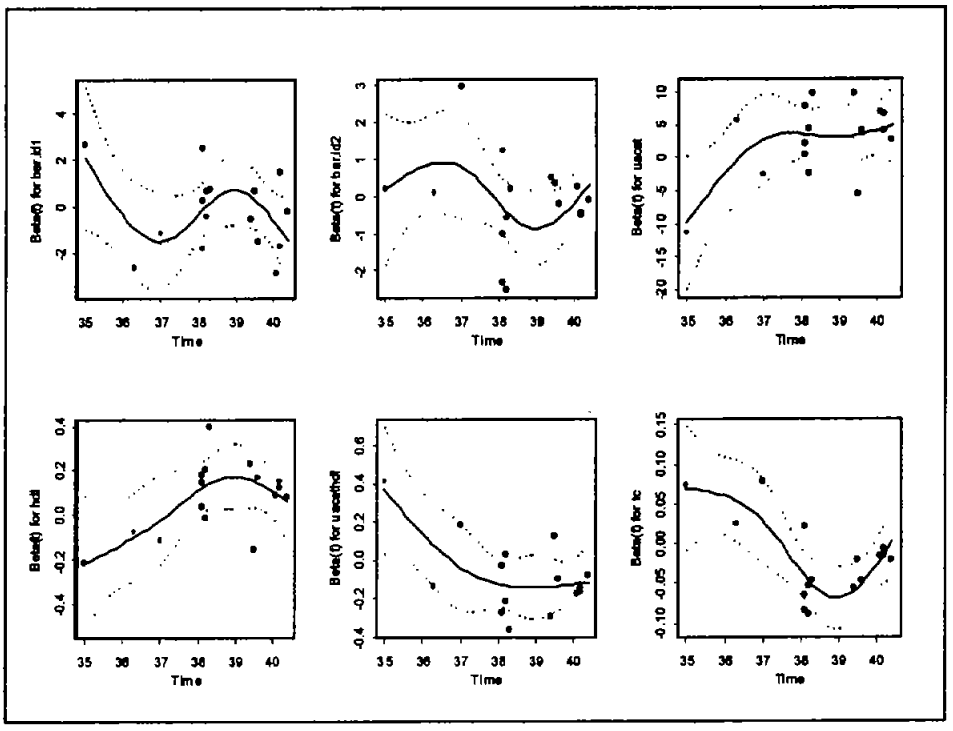

Figura 2.12. Gráficas de diagnóstico de la constancia de los coeficientes en el Modelo 3. Cada gráfica corresponde a un coeficiente del modelo contra el tiempo de sobrevivencia ordenado. Se muestra una línea de suavizamiento junto con las bandas de ± 2 desviaciones estándar.

En realidad este modelo no necesariamente debe de cumplir el supuesto de riesgos proporcionales porque las variables no cambian al mismo tiempo en todas las pacientes, ya que las visitas no se realizan para todas en las mismas semanas de embarazo.

El problema con este modelo (3) es que el número de las pacientes es de 16, que en realidad es un número muy pequeño para poder ajustar correctamente un modelo. Por lo tanto, por falta de datos, no se puede considerar este modelo como adecuado.

Debido a que tc es la única variable que no cumple el supuesto de proporcionalidad, se quitó del modelo par ver qué sucedía. Los coeficientes obtenidos no son significativamente diferentes a los del Modelo 3, pero el orden de los coeficientes de bar_id cambia, además de aumentar la significancia. Esto sugiere una posible interacción entre tc y el grupo, la cual podría ser posible debido al contenido de las barras. También se ajustó un modelo con la

interacción entre estas dos variables, la cual no resultó significativa ni afectó el comportamiento global del modelo. Sin embargo, si ajustamos la interacción sin el efecto principal de tc, el único término significativo es el del primer grupo, lo que podría sugerir que la ingesta de las barras podría afectar los niveles de colesterol.

CONCLUSIONES

De acuerdo a los modelos ajustados, en general se podría concluir lo siguiente:

Cuando los niveles de ácido úrico en las pacientes son altos, el riesgo de terminar prematuramente el embarazo se incrementa de manera importante. Lo mismo sucede con las lipoproteínas de alta densidad. Sin embargo, por alguna razón, cuando el nivel del ácido úrico es mayor al percentil 75, conforme aumenta el nivel de hdl el riesgo disminuye en gran medida. Al parecer también el colesterol total es una variable significativa y conforme va aumentando su nivel el riesgo de parto prematuro disminuye, pero no de manera importante.

Con respecto a los grupos, dado que en los modelos obtenidos la variable bar_id no resultó significativa, no se puede concluir que en alguno de los tres el riesgo de terminar prematuramente el embarazo sea mayor o menor. Sin embargo, al analizar los coeficientes de las interacciones entre grupo y diagnóstico en el Modelo 1, se observa que el riesgo de terminar prematuramente el embarazo para las pacientes que presentan preeclampsia es menor en el Grupo 2. Cabe aclarar que no existe ningún control sobre la ingesta de la barra y puede ser que, aún cuando una mujer esté en el grupo con suplemento vitamínico o de L-arginina, no se lo esté tomando.

Con respecto al colesterol total, los modelos obtenidos sugerían que posiblemente la ingesta de las barras podía aumentar el nivel de esta variable, pero no se pudo comprobar ya que con los datos que se tienen los coeficientes de la variable y las interacciones no eran significativos.

Mientras más tiempo duraron las pacientes dentro del estudio, menor era su riesgo de parto prematuro, lo que es razonable, ya que al realizar las visitas para revisión con regularidad tenían controlados la alimentación, peso, presión arterial, etc. Además, al estarse cuidando hubo muy poca incidencia de preeclampsia.

Sin embargo, no se puede saber si en realidad estas conclusiones son ciertas, ya que el número de pacientes utilizado es muy pequeño. Cabe aclarar que, aunque la base de datos hubiera sido lo suficientemente grande como para obtener algunas conclusiones certeras, éstas sólo serían válidas para el segmento de la población de mujeres embarazadas que asiste al Instituto Nacional de Perinatología, es decir, mujeres de un estrato social medio.

3.1 Limitaciones

Al realizar la presente tesis se tuvieron algunas limitaciones que no permitieron un mayor alcance al estudio.

Una limitación importante fue el tamaño de la base de datos y su contenido. Para empezar, el número de pacientes registradas (55) no era suficiente para que los modelos ajustados pudieran generalizarse a la población. Además, de esas pacientes menos de 46 tenían registrados los valores de todas las variables utilizadas y mencionadas en la sección 2.2, lo que hacía que para cada grupo de tratamiento el número de registros fuera muy pequeño y no se pudiera concluir si la L-arginina ayudaba a prevenir la preeclampsia. Además, muy pocas mujeres presentaron preeclampsia, quizás debido al cuidado que tuvieron durante el embarazo. Por otro lado, una paciente que estaba en el Grupo 3 informó que no se tomó el suplemento todo el tiempo porque las barras se ponían negras, esto, debido a que se oxidaban y, aunque esto no provocaba ningún problema, ni si quiera en el sabor, es posible que más mujeres hayan hecho lo mismo.

Dentro de las variables consideradas por el INPer para el estudio, se hacía un análisis en sangre del nivel de L-arginina en la paciente. Esta variable pudo ser importante, ya que tal sustancia puede obtenerse de los alimentos; además, hubiera ayudado en el análisis en el caso en que las pacientes no se comieran las barras o para considerar en vez de los grupos de tratamiento directamente el nivel de L-arginina en la sangre. Sin embargo, ésta y otras variables que indicaban, en caso de presentar preeclampsia, en qué número de visita había sido, no estaban registradas hasta el momento en el que se le proporcionó la base al IIMAS.

REFERENCIAS Y BIBLIOGRAFIA

- Collett, D. (1994) *Modelling survival data in medical research*, London: Chapman & Hall.

- Cox D.R. and Snell, E.J. (1968) "A general definition of residuals (with discussion)", Journal of the Royal Statistical Society, A, 30, 248-75.

- Cox, D.R. (1972) "Regression models and life tables (with discussion)", Journal of the Royal Statistical Society, B, 74, 197-220.

- Crowley, J. and Hu, M. (1977) "Covariance analysis of heart transplant survival data", *Journal of the American Statistical Association*, 72, 27-36.

- Fleming, T. R. and Harrington, D. P. (1991) Counting processes and Survival Analysis, New York: Wiley.

- Peto, R., et. al. (1977) "Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples", British Journal of Cancer, 34, 57-67.

- Pettitt, A.N. and Bin Daud, I. (1989) "Case-weighted measures of influence for proportional hazards regression", Applied Statistics, 38, 51-67.

- Schoenfeld, D.A. (1982) "Partial residuals for the proportional hazards regression model", Biometrika, 69, 239-41.

- Tableman, M. & Kim, J. S. (2004) Survival Analysis using S: Analysis of time-to-event data, Boca Raton, Florida : Chapman and Hall/CRC.

- Therneau, T.M., Granbsch, P.M. and Fleming, T.R. (1990) "Martingale-based residuals for survival models", Biometrika, 77, 147-60.