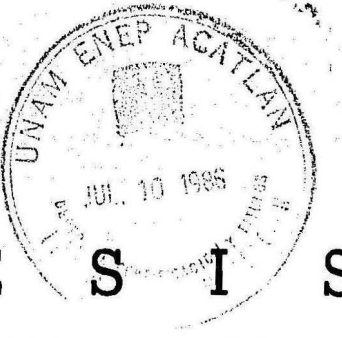
Universidad Nacional Autónoma de México

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES

''ACATLAN''

AJUSTE DE MODELOS PARA DATOS DE SUPERVIVENCIA EN GLIM



T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I A

P R E S E N T A :

Leticia del Carmen de la Cruz Mejía

ACATLAN, EDO. DE MEX. 1986

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N D I C E

INTRODUCCION

El análisis estadístico de la duración de la vida, ha sido analizado desde hace mucho tiempo, en particular por los actuarios. Un interés reciente surge de examinar el efecto que tienen posibles variables explicativas sobre la duración de la vida o sobre el fenómeno de supervivencia. Aunque existe una amplia bibliografía que aborda el problema, se hace necesaria una revisión de los enfoques utilizados para su análisis, así también una breve discusión de algunos métodos de ajuste de modelos; tal es el objetivo de esta tesis.

Dentro del análisis de supervivencia con variables explicativas, uno de los problemas mas importantes es la estimación de los parámetros relacionados con dichas variables, introducidas en el modelo propuesto, para un problema particular.

El problema relacionado con la estimación de los parámetros depende principalmente del modelo que se proponga y de la forma en que los datos quedan involucrados con respecto al tipo de mecanismo de censura; ya que de este depende, principalmente, la complejidad que pueda surgir en el cálculo de los estimadores.

Para esto, existen ya elaborados métodos para el cálculo de los estimadores y la inferencia estadística involucrada (ver Kalbfleisch y Prentice, 1980, cap. 3), para algunos modelos particulares.

Por otro lado, para llevar a cabo la estimación, así como el análisis estadístico asociado, se hace necesario la elaboración de un programa de cómputo por medio del cual se lleve a cabo el ajuste.

En este sentido, existe un paquete llamado GLIM (Generalized Linear Interactive Modelling), que sirve para el ajuste de modelos lineales generalizados. Este paquete tiene amplias ventajas para el análisis del problema. Se plantea entonces el problema de trasladar nuestros modelos propuestos al contexto de modelos lineales generalizados, por esto se hace necesario considerar las características básicas de estos.

Los modelos que se proponen para el análisis de datos de supervivencia podemos dividirlos, para propósitos de mostrar su adaptación en GLIM, en dos clases: modelos paramétricos y modelos semiparamétricos.

Los modelos paramétricos, son aquellos en los cuales se propone la función de riesgo explícitamente, es decir, se expresa totalmente la forma de la distribución de los tiempos de falla. Para hacer esto seguiremos el enfoque de

Aitkin y Clayton (1980) y posteriormente el de Roger y Peacock (1980).

Los modelos semi-paramétricos, son aquellos en los cuales solo se describe parcialmente la distribución del tiempo de falla, siendo la otra parte arbitraria. Ejemplo de este tipo de modelo es el que propone Cox (1972) del cual abordaremos el problema de implantación con el paquete GLIM siguiendo el enfoque de Whitehead (1980).

Para el desarrollo del tema en cuestión, se ha considerado conveniente estructurar este trabajo en cuatro capítulos: el capítulo I contiene la terminología básica relacionada con el problema, para poder hacer mas fácil la lectura de esta tesis.

En el capítulo II presentaremos los modelos lineales generalizados. En el III presentaremos como los datos de supervivencia se pueden adaptar a un modelo lineal generalizado, y en el último capítulo mostraremos la construcción de la función de supervivencia para los modelos considerados, así como la bondad de ajuste de estos.

Cabe aclarar que en algunos parrafos aparecen ciertos números, entre paréntesis, que indican una nota aclaratoria; en la parte final del trabajo.

C A P I T U L O I

CONCEPTOS PRELIMINARES

1.1 Tiempos de Falla

En el análisis de datos de supervivencia, el interés se centra en un grupo de individuos u objetos, para los cuales hay definido un evento: muerte,descompostura, enfermedad,etc. el cual llamaremos "falla".

Al tiempo que transcurre, hasta la ocurrencia de ese evento se le llama "tiempo de falla". Como podemos observar, la variable aleatoria determinante es el tiempo de falla,al cual denotaremos por T.

Existen diferentes casos que pueden presentarse en el proceso de supervivencia, tomando en cuenta el tiempo de falla que tienen asociados dentro del proceso.

Supongamos que tenemos un conjunto de M individuos [1] bajo estudio; aquí se pueden presentar dos casos:

a) Que todos los individuos inicien el estudio en un mismo tiempo y cada individuo tenga un tiempo de falla, t_1 t_2 ...$t_M$respectivamente con $t_i \neq t_j$ $\forall$ $i \neq j$; $i,j=1...M$ (no existen fallas múltiples).

b) Que los individuos presenten tiempos diferentes de entrada al estudio. En este caso se tendrá que considerar el tiempo de seguimiento, que es el tiempo total desde que un individuo entra en el estudio hasta que falla.

Tanto en a) como en b) se puede presentar que los tiempos de falla coincidan, es decir, si t_1 ...t_M son los tiempos de falla correspondientes, se puede tener $t_i = t_j$ para algún $i,j=1,...M$ y, en este caso, se dirá que se tienen fallas múltiples.

Evidentemente el que se tengan tiempos de supervivencia iguales depende de las unidades de tiempo que se esten manejando (días,semanas,meses,etc.), lo cual puede evitarse reduciendo o afinando la escala de medición del tiempo asignada al fenómeno y asi evitar las fallas múltiples. Este no es el caso cuando se ha realizado el estudio en ciertas condiciones y ya no es posible modificar los tiempos asignados.

1.2 Censuras

Una fuente principal de dificultad en el análisis de datos de supervivencia, es la posibilidad de que algunos individuos no puedan ser observados para todo el tiempo de falla, es decir, que por diversas causas desconocidas, un conjunto de r individuos abandonen el estudio en tiempos $t_1, t_1 \ldots t_r$ (a estos tiempos les llamaremos tiempos de censura); al igual que en los tiempos de falla, aquí pueden presentarse tiempos de censura múltiple.

En un contexto general, de acuerdo al tipo de estudio que se lleve a cabo, es posible clasificar los tiempos de censura, en dos tipos: censura por la derecha y censuras por la izquierda. Las censuras por la derecha, son aquellas en las que, a partir de un tiempo determinado, la variable de interés ya no puede ser observada. Las censuras por la izquierda suceden cuando existe la posibilidad de que algunas fallas hayan sucedido antes del inicio de nuestro estudio y por lo tanto no sean registradas.

ejemplo:
Un investigador quiere conocer la edad a la cual cierto grupo de niños aprende a ejecutar una tarea particular. Cuando este llega al lugar de estudio, encuentra que algunos niños, ya saben como realizar cierto tipo de tarea; estos contribuyen a tiempos de censura por la izquierda. Algunos niños aprenden a realizar esta, mientras él permanece en el lugar (tiempos de falla). Cuando el investigador se va, algunos niños todavía no han aprendido a realizar la tarea, estos contribuyen a la censura por la derecha.
En todo el análisis posterior, sólo consideraremos las censuras por la derecha.

Los tiempos de censura por la derecha pueden plantearse de dos formas diferentes, a saber :
a) Censura tipo I. Si deseamos terminar el estudio a un tiempo previamente asignado.
b) Censura tipo II. Si deseamos terminar el estudio, cuando un número preasignado de fallas ocurran.
El problema del análisis de datos censurados, plantea la cuestión de cómo deben ser considerados estos, con respecto a sus tiempos de supervivencia; ya que no se conocen los tiempos de falla de cada individuo censurado y, por tanto, no es posible, saber cual es su distribución de tiempo de falla correspondiente.

1.3 El Problema

El fenómeno de supervivencia plantea dos problemas, a saber:

a) Si se considera un conjunto de individuos sujetos a un proceso de supervivencia, con datos censurados y no censurados, podemos plantear la cuestión: ¿Cuál es la forma funcional que describe el proceso de supervivencia para todo tiempo ?

b) Si en correspondencia con el fenómeno considerado, se presentan un conjunto de variables, que influyen o pueden influir en el comportamiento de la función de supervivencia, (llamadas variables explicativas, concomitantes o covariables denotadas con $Z_1, Z_2, \ldots, Z_K$). ¿Cuáles son los pesos relativos de cada variable y por tanto cuales son las variables que influyen determinantemente sobre la función de supervivencia ?.

Una parte importante en el análisis del fenómeno de supervivencia, es el de la construcción empírica de la función de supervivencia [2]. Sin embargo, no será considerada en este trabajo.

A continuación, plantearemos la terminología de la descripción matemática de los fenómenos de supervivencia.

1.4 Distribución de Tiempo de Falla

Consideremos la variable aleatoria determinante del fenómeno de supervivencia T, la cual denota el tiempo de falla . La función de distribución acumulativa:

$$F_T(t) = \text{Prob}\{ T \leq t \}$$

es llamada distribución de tiempo de vida o distribución de tiempo de falla. Asociada con esta función se define la función:

$$S_T(t) = 1 - F_T(t)$$

$$S_T(t) = \text{Prob}\{ T > t \}$$

Esta es la función de distribución de supervivencia o simplemente, función de supervivencia. En un fenómeno de supervivencia, el tiempo siempre es positivo por tanto se debe cumplir la relación :

$$S_T(0) = 1 \quad y \quad F_T(0) = 0$$

La función de densidad de probabilidad, asociada con F es la función f (t), definida como:

$$f(t)=\frac{dF_T(t)}{dt},$$

esta función es llamada también curva de muerte en el análisis de supervivencia y, por definición, es una tasa instantánea absoluta de muerte.

Otra función de utilidad, es la función de riesgo de muerte o tasa de riesgo de falla, denotada por $\lambda(t)$ [3] y definida como:

$$\lambda_T(t) = -\frac{f_T(t)}{S_T(t)},$$

de la definición de $S_T(t)$, se obtiene la relación:

$$\lambda_T(t)=-\frac{d}{dt}(\ln S_T(t))$$

finalmente, se puede definir la función acumulativa de riesgo como:

$$\Lambda_T(t)=\int_0^t \lambda(x)dx=-\ln S_T(t).$$

La función de supervivencia se puede expresar en términos de $\lambda(t)$ como:

$$S_T(t)=\exp\left\{-\int_0^t \lambda_T(u)\,du\right\}=\exp\left\{-\Lambda_T(t)\right\}.$$

1.5 Planteamiento de los modelos

En esta sección, revisaremos algunos de los posibles modelos que pueden ser usados para representar el efecto, sobre el tiempo de falla, de las variables explicativas. Para esto, supondremos que, para cada individuo, hay definido un vector Z de variables explicativas de la forma:

$$Z=(Z_1, Z_2, \ldots, Z_K) \qquad \text{(k entradas)},$$

que para cada individuo (j), da lugar a un vector Z de variables expicativas

$$Z_j=(Z_{1j}, Z_{2j}, \ldots Z_{Kj}).$$

Es prudente aclarar que las componentes de Z pueden ser de diferentes tipos, en función de la característica a que estan asociadas y, por ende a la manera en cómo afectan al tiempo de falla y pueden agruparse en:

i) tratamiento,
ii) propiedades intrínsecas del individuo,
iii) variables exógenas,

Algunos ejemplos de los tipos de variables explicativas, son los siguientes:

En una comparación simple de dos tratamientos; por ejemplo, de un tratamiento nuevo con uno que sirva de control, se considera una variable aleatoria explicativa binaria con valores: igual a uno para individuos que reciben el tratamiento e igual a cero para los individuos que reciben el de control. Si el tratamiento es especificado por el nivel de dósis la correspondiente variable explicativa es dósis.

Las variables explicativas, que miden las propiedades intrínsecas de los individuos incluyen, en el contexto médico, variables demográficas tales como: sexo, edad, y variables que describen la historia médica antes de admisión al estudio,etc. Estas variables pueden dar lugar a agrupaciones cualitativas de los individuos.

Finalmente, las variables exógenas definen, en particular, características del medio ambiente de el problema.

Otra manera, en la cual se pueden clasificar las variables explicativas es en, constantes o dependientes del tiempo. Para efectos de este trabajo y por cuestión de simplicidad supondremos en principio que Z no es función de t.

Pasemos ahora a tratar de resolver nuestro problema que planteamos al principio y, para fijar ideas, supongamos que tenemos conocida la forma funcional de S(t), dada por:

$$S(t)=\exp\left\{\int_0^t \lambda(t)dt\right\},$$

donde ya se ha determinado $\lambda(t)$. Esta función de riesgo $\lambda(t)$, es la que nos será de utilidad para proponer un modelo explicativo, con base en nuestra información, acerca de las variables que consideremos más determinantes del fenómeno de supervivencia. Una manera de introducir la información de esas variables explicativas, es en la función $\lambda(t)$ y precisamente como función de las variables explicativas. Nuestro modelo se reducirá, a proponer una forma explícita funcional para $\lambda(t)$, en términos de las covariables. Dos posibles modelos son:

$$\lambda(t,Z,\beta)=\lambda_0(t)+g(Z,\beta)\ , \qquad (1.1)$$

$$\lambda(t,Z,\beta)=\lambda_0(t)g(Z,\beta)\ . \qquad (1.2)$$

donde $\lambda_j(t)$, es cierta función no especificada del tiempo, llamada función fundamental y, $g(Z,\beta)$, es una función general de las variables explicativas, las cuales se han, arreglado, de tal manera, que formen un vector $Z_j=(Z_{1j},\ldots Z_{Kj})$ y, un vector de coeficientes desconocidos $\beta=(\beta_1,\ldots \beta_K)$, asociados con las variables concomitantes Z, donde cada j=1,...N, corresponde a los individuos que estan en el estudio.

Al modelo 1.1 se le llama modelo aditivo, Aranda-Ordaz (1983), y al modelo 1.2 modelo multiplicativo por razones obvias. En lo sucesivo solamente utilizaremos el modelo 1.2.

Algunas proposiciones para la función $g(Z,\beta)$, en el modelo 1.2 son:

$$\text{i)}\quad g(Z,\beta)=\frac{1}{1+Z\cdot\beta}\quad ,$$

$$\text{ii)}\quad g(Z,\beta)=\exp(Z\cdot\beta)\quad ;$$

donde $Z\cdot\beta$ es el producto escalar entre los vectores $(Z_1,\ldots Z_K)$ y $(\beta_1,\ldots\beta_K)$, en lo cual si $Z=0$ $g(Z,\beta)=1$, es decir, en ausencia de covariables el modelo corresponde a

$$\lambda(t,Z,\beta)=\lambda_0(t)\ .$$

Los modelos anteriores para $g(Z,\beta)$ han sido ampliamente estudiados; el modelo ii) tiene especial relevancia dado que es el más utilizado al abordar problemas de supervivencia.

El modelo 1.2 junto con la proposicion ii) nos dan el modelo siguiente que es el que será estudiado a lo largo de este trabajo.[4]

$$\lambda(t,Z,\beta)=\lambda_0(t)\exp(Z\cdot\beta)\ .$$

Cuando la función fundamental $\lambda_0(t)$ se propone explícitamente, salvo estimaciones de algunos parámetros, se dice que se esta proponiendo un modelo paramétrico de análisis. En el caso en que se deja la funcion $\lambda_0(t)$ "arbitraria" diremos que se propone un modelo semiparamétrico de análisis.

1.6 Construcción de la función de verosimilitud

En seguida deseamos obtener en general, suponiendo arbitraria $\lambda_0(t)$, la mejor estimación de los valores de los parametros β y, para esto usaremos el método de máxima verosimilitud.

Sea Z un vector de k variables concomitantes y, consideremos la función de riesgo $\lambda(t,Z,\beta)$ que por brevedad denotaremos como $\lambda(t,Z)$. Asi también la función acumulativa $\Lambda(t,z)$ esta dada por:

$$\Lambda(t,Z)=\int_0^t \lambda(u,Z)du$$

y la función de supervivencia

$$S(t,Z) = \exp\,[\,-\Lambda(t,Z)\,]\,.$$

En términos de S(t) y $\lambda(t)$, obtenemos la función de densidad como:

$$f(t,Z) = \lambda(t,Z) \times S(t,Z)\,.$$

Si Z_j representa el vector de covariables observadas del individuo (j), y que son independientes del tiempo; para cada individuo (j) podemos definir una función de riesgo $\lambda(t,Z_j)$ y función de supervivencia $S(t,Z_j)$.

Consideremos una muestra de M individuos que participan en algun tiempo en el estudio sea τ_j el tiempo en el cual el individuo (j) entra en el estudio y t_j el tiempo en el cual (j) fue observado por última vez.

Denotaremos por Ω_0 el conjunto de d individuos muertos y por Ω el conjunto de (M-d) individuos que estuvieron vivos hasta que fueron observados por última vez en el estudio completo. Bajo estas consideraciones la función de verosimilitud correspondiente será:

$$L = \prod_{j\in\Omega_0}\left[\frac{f(t_j,Z_j)}{S(\tau_j,Z_j)}\right] \times \prod_{j\in\Omega}\left[\frac{S(t_\ell,Z_\ell)}{S(\tau_\ell,Z_\ell)}\right]$$

o,

$$L=\prod_{j\in\Omega_0}\left[\lambda(t_j,Z_j)\right] \times \prod_{j\in\Omega}\left[\frac{S(t_j,Z_j)}{S(\tau_j,Z_j)}\right]\,. \qquad (1.3)$$

Si definimos la variable indicadora

$$\delta_j = \begin{cases} 1 \text{ si } (j) \text{ muere al tiempo } t_j, \\ 0 \text{ si } (j) \text{ estuvo vivo al tiempo } t_j, \end{cases}$$

Entonces la ecuación 1.3 se puede escribir

$$L=\prod_{j=1}^{M}\left[\lambda(t_j,Z_j)\right]^{\delta_j}\times\left[\frac{S(t_j,Z_j)}{S(\tau_j,Z_j)}\right]. \qquad (1.4)$$

Si consideramos el caso particular en el que los M individuos entran a t=0, entonces =0 y S(,Z)=1 para todo j, en tal situación

$$L=\prod_{j=1}^{M}\left[\lambda(t_j,Z_j)\right]^{\delta_j}\times\left[S(t_j,Z_j)\right]. \qquad (1.5)$$

Una vez que se propone de alguna manera la función $\lambda(t,Z)$ se puede hallar L introduciendo esta función en la ecuación 1.5 .

Considerando el modelo multiplicativo propuesto:

$$\lambda(t,Z)=\lambda_0(t)\exp(Z\beta),$$

la verosimilitud será:

$$L=\prod_{j=1}^{M}\left[\lambda_0(t_j)\exp(Z_j\cdot\beta)\right]^{\delta_j}\times\left[S(t_j,Z_j)\right],$$

donde

$$S(t_j,Z_j)=\exp\left\{-\int_0^{t_j}\lambda_0(t)\exp(Z_j\beta)\,dt\right\},$$

simplificando se obtiene

$$L=\prod_{j=1}^{M}\left[\lambda_0(t_j)\right]^{\delta_j}\exp\left\{\delta_j Z_j\beta-\exp(Z_j\beta)\cdot\int_0^{t_j}\lambda_0(u)\,du\right\}, \qquad (1.6)$$

si tomamos en cuenta que en $\lambda(t)$ aparecen s parámetros desconocidos, entonces las ecuaciones de verosimilitud serán:

$$\frac{\partial L}{\partial \alpha}=0 \quad ; \qquad (j=1,\ldots s),$$

$$\frac{\partial L}{\partial \beta}=0 \quad ; \qquad (j=0,1,\ldots k).$$

o bien si tomamos $\Xi = \ln L$, las ecuaciones quedan:

$$\frac{\partial \Xi}{\partial \alpha} = 0 \quad ; \qquad (j=1,\ldots s),$$

$$\frac{\partial \Xi}{\partial \beta} = 0 \quad ; \qquad (j=0,1\ldots k).$$

Al analizar las ecuaciones de verosimilitud 1.6 observamos que tenemos practicamente dos problemas de estimación; uno es la estimación de los parámetros β_j y otro, la estimación de $\lambda_0(t)$, que tambien involucra parámetros α_j desconocidos. Existen básicamente dos tipos de enfoque para la estimación de $\lambda_0(t)$ y de los parámetros .

El primero consiste en proponer una forma explicita para $\lambda_0(t)$ que involucre ciertos parámetros a determinar junto con los parámetros β_j involucrados (modelo paramétrico). El segundo es el "modelo de riesgos proporcionales", en el cual se construye la función de "verosimilitud parcial" en la cual se elimina la función fundamental $\lambda_0(t)$; solamente se supone $\lambda_0(t) > 0$ y contínua, obteniendo así, la estimación del vector β_j . Una vez calculado β se puede encontrar una estimacion para la funcion $\lambda_0(t)$ y de aqui obtener una estimación de S (t).

Para efectuar la estimación de los parámetros involucrados en los casos paramétricos y semiparamétricos, se utilizarán modelos lineales generalizados pues estos son suceptibles de ser ajustados mediante el paquete GLIM, que se encuentra disponible en la UNAM.

CAPITULO II

MODELOS LINEALES GENERALIZADOS

Para realizar la estimación de los parámetros involucrados en el caso Paramétrico y Semiparamétrico respectivamente, se presenta un cambio de enfoque el cual consiste en trasladar los modelos planteados al contexto de los Modelos Lineales Generalizados y posteriormente su implantación mediante el uso del paquete GLIM (Generalized Linear Interactive Modelling).

Dado que para uso del paquete GLIM es fundamental el concepto de modelo lineal generalizado, presentaremos las ideas básicas involucradas en la definición de este último.

Los modelos Lineales Generalizados (MLG) son una extensión de los modelos lineales clásicos. Para un modelo lineal clásico se tiene una parte sistemática y una parte aleatoria; la parte sistemática se refiere fundamentalmente a la parte lineal y la aleatoria a la estructura asociada del error.

2.1 Parte Sistemática del Modelo

Sea el vector de observaciones $y = (y_1, y_2, \ldots y_M)$, de variable aleatoria Y que se distribuyen independientemente con media μ, y varianza constante. Y supongamos la existencia de covariables $Z_1, Z_2, \ldots Z_K$ con valores conocidos, tales que μ se puede expresar como una combinación lineal de las Z, o sea:

$$\mu = \sum_{j=1}^{K} Z_j \beta_j \quad ,$$

donde las β_j son parámetros desconocidos y deben ser estimados de los datos. En términos de cada observación queda:

$$E(Y_i) = \mu_i = \sum_{j=1}^{K} Z_{ij} \beta_j \quad ,$$

donde $i = 1, 2, \ldots M$ y Z_{ij} es el valor de la j-ésima covariable para la i-ésima observación. En notación matricial:

$$\mu = Z\beta \quad ,$$

donde μ es una vector de Mx1, Z una matriz de Mxk y β de kx1. Esto completa la especificación de la parte sistemática del modelo.

Así, el modelo clásico es aquel que cumple con tener $Y_1, Y_2, \ldots, Y_n$ variables con distribución normal y varianza constante, y además:

$$E(Y) = \mu, \qquad \mu = Z \cdot \beta .$$

Para generalizar, las condiciones anteriores se modifican ligeramente de la siguiente manera:

Primero.- Se tiene una compomente aleatoria formada por un vector de Y variables independientes con parámetro cuya distrribución pertenece a la familia exponencial.

Segundo.- Una componente sistemática que involucra covariables $Z_1, \ldots, Z_K$ y producen un "predictor lineal" dado por:

$$\eta = \sum_{j=1}^{K} Z_j \beta_j .$$

Tercero.-una función monótona diferenciable que relaciona la parte sistemática con la aleatoria, llamada "funcion liga"

$$\eta_i = g(\mu_i) .$$

En suma, un modelo lineal generalizado estará caracterizado por los tres componentes anteriores, donde la distribución del error pertenece a la familia exponencial, la parte sistemática se mantiene lineal y la funcion liga debe ser monótona diferenciable.

2.2 Parte aleatoria del modelo

La familia exponencial puede caracterizarse como sigue:

$$f(Y,\theta,\phi) = \exp\left\{ \alpha(\phi)\left[\theta Y - b(\theta) + h(Y)\right] + c(\phi, Y) \right\},$$

donde $\alpha(\phi), b(\theta), c(Y,\phi)$ son funciones especificas;con ϕ conocida, la media y la varianza se obtienen mediante las relaciones

$$E(Y) = \mu = b'(\theta) ,$$

$$V(Y) = b''(\theta)\alpha(\phi),$$

en la cual (') denota la derivada con respecto al parámetro y (''), la segunda derivada respectiva.

a la cantidad θ se la llama parámetro canónico de la familia exponencial y a ϕ se la llama parámetro de dispersión y $b''(\theta)$ recibe el nombre de función de varianza. Comunmente la función $a(\phi)$ es de la forma:

$$a(\phi)=\frac{\phi}{\omega} \quad ,$$

donde W es cierta "función de peso"

2.3 Función Liga

La función liga relaciona el "predictor lineal" η con el valor del parámetro θ del dato y. Cuando se tiene un modelo lineal clasico μ y η son iguales y el modelo tendra asociada la función liga identidad, en la cual μ y η pueden tomar cualquier valor real.

Sin embargo en algunos casos es necesario que μ sea positivo y no conviene usar la liga identidad.[5]

Las funciones liga que son usadas de acuerdo al tipo de distribucion que se considere,son:

Normal $\quad \eta = \mu \quad ,$

Poisson $\quad \eta = \ln \mu \quad ,$

Binomial $\quad \eta = \ln\left[\frac{\mu}{1-\mu}\right] \quad ,$

Gamma $\quad \eta = \mu^{-1} \quad ,$

Gaussiana Inversa $\quad \eta = \mu^{-2} \quad ,$

las anteriores son llamadas ligas canónicas.

2.4 Estimación de los parámetros en la expresión lineal

Para la estimación de los parámetros β_j y de allí el predictor η_i y los valores ajustados μ_i , se usa el método de máxima verosimilitud.

Para hacerlo, se supone que la matriz de covariables Z_{ij} es de rango completo [6]. Considerando la verosimilitud como una función de β y maximizando para β

se puede obtener una ecuacion de la forma:

$$A\beta = r \quad ,$$

donde la matriz A y el vector r son, en general, función de μ el cual es desconocido debido a que es función de β por tanto no es posible hallar una solución directa. Sin embargo puede resolverse iterativamente: primero tomamos las observaciones como estimadores iniciales de μ , resolvemos para β usamos esos valores de β para calcular los nuevos estimadores de μ y de nuevo se resuelven para β y asi sucesivamente hasta la convergencia de β a $\hat{\beta}$, Nelder y Wedderburn (1972) muestran que la solución a las ecuaciones de m.v. es equivalente a un proceso iterativo de mínimos cuadrados ponderados con función de peso:

$$\omega = \left(\frac{d\mu}{d\eta}\right)^2 \Big/ V \quad ,$$

en la cual μ es la media y $\eta = \sum_i Z_i \beta_i$

y se tiene una variable respuesta modificada

$$y = \eta + (Y - \mu)\frac{d\mu}{d\eta} \quad .$$

Así, un proceso de Newton Raphson con segundas derivadas para una muestra de tamaño M da:

$$A\delta\beta = C \quad , \qquad (2.1)$$

A es una matriz de tamaño mxm, dada por:

$$A_{ij} = \sum \omega_l Z_{il} Z_{jl} \quad ,$$

C es un vector de dimensión mx1 de la forma:

$$C_i = \sum \omega_l Z_{il} (Y - \mu)\frac{d\mu}{d\eta} \quad ,$$

con $$\sum_j A_{ij}\beta_j = \sum_l \omega_l Z_{il}\eta_l \quad .$$

La ecuación 2.1 se puede escribir como:

$$A\beta^* = r \quad ,$$

donde $$r_i = \sum \omega_l Z_{il} y_l \quad ; \quad y_l = \eta_l + (Y_l - \mu_l)\frac{d\mu_l}{d\eta_l}$$

y $$\beta^* = \beta + \delta\beta \quad .$$

2.5 Bondad de Ajuste

Podemos preguntarnos ¿qué tan bien apoyan los datos al modelo propuesto?. Suponiendo que el error y la función liga son buenos, consideremos la estructura lineal para tratar de responder la pregunta.

La estructura lineal es la suma de los efectos de las variables concomitantes y su composición expresa la influencia de esas variables sobre la variable dependiente. Los datos dan informacion acerca de cuales efectos tienen influencia importante y cuales no. Así nuestro propósito es obtener la mejor elección entre el número de variables y sus parámetros que debieron ser incluidos en la estructura lineal y la capacidad del modelo para representar los datos.

Podemos distinguir cinco formas especiales para la estructura lineal como sigue:

Si se incluyen n parámetros linealmente independientes en la estructura lineal del estimador de m.v. para β será equivalente a las observaciones mismas. Este modelo es conocido como "modelo completo" o modelo saturado; los datos son reproducidos exactamente, pero sin ninguna simplificación en la interpretación. En cambio, si se propone un valor común para las μ , se tendrá un "modelo nulo"; este es un modelo simple pero comunmente no representa adecuadamente la estructura de los datos.

Existen también otros modelos menos extremos, uno es el "modelo mínimo" que considera a aquellos parámetros que deben estar en el modelo. El modelo mas grande es el"modelo máximo", y entre esos dos extremos esta el modelo que estamos considerando el "modelo corriente".

Nuestro problema es determinar la utilidad de un parámetro para el modelo corriente o de otro modo, el error de ajuste inducido por la inclusión del parámetro.

Nosotros consideramos la aceptabilidad del modelo corriente relativo al modelo completo,comparando la verosimilitud del modelo corriente (lc) con la del modelo completo (lf) para los datos dados, asi se considera la estadística:

$$S(c,f)=-2\log(lc/lf)\,, \qquad (2.2)$$

donde S es llamada "desviación escalada" si usamos para lc y lf, la estructura del error que consideramos en 2.2 con los estimadores m.v. de los parámetros, se obtiene:

$$S(c,f)=2\sum\{y(\hat{\theta}-\tilde{\theta}) + b(\hat{\theta})-b(\tilde{\theta}) / a(\phi)\}\,,$$

$\hat{\theta}_j$ y $\tilde{\theta}_j$ son los estimadores m.v. de θ_j bajo el modelo corriente y completo respectivamente, escribiendo S(c,f) en términos de $\hat{\mu}$ y μ queda:

$$S(c,f)=\frac{D(c,f)}{\phi} \quad ; \quad a(\phi)=\frac{\phi}{w} .$$

D(C,f) es la desviación o "devianza" del modelo corriente relativo al modelo completo, es el parámetro de escala, la devianza es una cantidad conunmente usada en estadística; para distribuciones especiales se tiene:

devianza NORMAL $= \sum_i \left\{\frac{Y-\mu}{\sigma}\right\}^2$,

devianza POISSON $= 2\left[\sum_i \left\{Y\ln\left(\frac{Y}{\mu}\right) - (Y-\mu)\right\}\right]$,

devianza BINOMIAL $= 2\sum_i \left\{Y\ln\left(\frac{Y}{\mu}\right)+(n-Y)\ln\left(\frac{n-Y}{n-\mu}\right)\right\}$,

devianza GAMMA $= -2\sum_i \left\{\ln\left(\frac{Y}{\mu}\right) + \left(\frac{Y-\mu}{\mu}\right)\right\}$,

devianza GAUSSIANA INVERSA $\sum_i \left\{\frac{Y-\mu}{\mu}\right\}^2 \quad ; (i = 1,\ldots, M.)$.

2.6 Residuales

Una vez que se tiene ajustado un modelo lineal generalizado especificado por su varianza, su función liga y las covariables en el predictor lineal, entonces pueden plantearse las siguientes preguntas:

1.- ¿Son necesarias mas covariables en el predictor lineal?

2.- ¿Debe cambiarse alguna covariable Z por alguna función de Z particular por ejemplo: g(Z)=lnZ ?

3.- ¿Debe cambiarse la función liga por ejemplo de $\eta = \mu$ a $\eta = \ln\mu$?

4.- ¿Debe modificarse la función varianza por ejemplo de $V(\mu)=\lambda\mu$ a $v(\mu)=\lambda\mu^2$?

Para modelos normales, podemos expresar la variable respuesta como:

$$y=\mu+(y-\mu) ,$$

o sea: dato=valor ajustado + "residuo"

Los residuos pueden ser usados para explorar que tan adecuado es el ajuste de un modelo.

Para los M.L.G. es necesario generalizar el concepto de residual aplicables a todas las distribuciones que reemplazan a la normal y, además, que pueden ser usados para los mismos propósitos que los residuales normales.

Para esto existen tres formas de residuales generalizados, los cuales involucran μ en vez de $\hat{\mu}$, los cuales son los residuales de "Pearson" (r_P), "Ascombe"(r_A "residuales de desviación" (r_D).

$$r_P = \frac{y-\mu}{\sqrt{V(\mu)}},$$

$$r_A = \frac{3(y^{\frac{1}{3}}-\mu^{\frac{1}{3}})}{2\mu^{\frac{1}{3}}}, \qquad \text{(para Gama)}$$

$$r_A = \frac{3(y^{\frac{2}{3}}-\mu^{\frac{2}{3}})}{\mu^{\frac{1}{6}}}, \qquad \text{(para Poisson)}$$

$$r_D = sgn\left\{(y-\mu)[2(y\ln(\frac{y}{\mu})-y+\mu)]\right\}.$$

Los residuales Ascombe y residual de desviación son aparentemente diferentes pero dan,para cierto rango de valores,valores muy similares.

En GLIM se definen los residuales "estandarizados" como:

$$\text{residual} = (\text{observada} - \text{ajustada})\sqrt{\frac{(\text{función de peso})}{K\cdot\text{varianza}}},$$

K=parámetro de escala

Una gráfica que resulta de utilidad, después que se ha efectuado el ajuste, es la de los residuales contra valores ajustados (o alguna transformacion de ellos).Tal ajuste es capaz de revelar puntos aislados con residuales grandes, o una curvatura general, indicando escala de covariables inadecuado o función liga, o una tendencia en la dispersión con el aumento de los valores ajustados, indicando una función de varianza insatisfactoria.

CAPITULO III

AJUSTE DE MODELOS PARA DATOS DE SUPERVIVENCIA EN GLIM

3.1 Modelo paramétrico

Consideraremos $t_1, t_2, \ldots, t_r \ldots t_M$ tiempos de falla de M individuos de los cuales (M-r) son censurados, asociados con estos tiempos consideraremos los correspondientes valores de las variables explicativas Z_{ij} en donde i=1,2,..., M; j=0,1,...,k (con $Z_{i0}=1$). La función de densidad de falla f(t), la función de distribución F(t) y la función de riesgo $\lambda(t)$ estan relacionadas por la ecuación:

$$\lambda(t)=\frac{f(t)}{S(t)} \quad , \qquad (3.1)$$

donde

$$S(t)=1-F(t) \quad ,$$

la función de riesgo se propone de tal manera que involucre las variables explicativas a través de un modelo Log-lineal (modelo de riesgos proporcionales).

$$\lambda(t_i)=\lambda_0(t_i)\exp\left(\sum_{j=0}^{K} Z_{ij}\beta_j\right) . \qquad (3.2)$$

Así la función de densidad queda:

$$f(t_i)=\lambda(t_i)S(t_i) \quad ,$$

$$S(t_i)=\exp\left\{-\Lambda(t_i)\exp(Z_i\beta)\right\} \quad , \qquad (3.3)$$

donde

$$\Lambda(t_i)=\int_0^{t_i}\lambda(t)\,dt$$

$$f(t_i)=\lambda_0(t_i)\exp(Z_i\beta)\exp\left\{-\Lambda(t_i)\exp(Z_i\beta)\right\} . \qquad (3.4)$$

En seguida construiremos la función de verosimilitud. Para esto consideraremos la variable indicadora δ_i que toma el valor de 1 para el dato no censurado y 0 para censura. Además supondremos que el mecanismo de censura es independiente de la variable explicativa. Así la función de verosimilitud queda:

$$L = \prod_{i=1}^{M} \left[f(t_i) \right]^{\delta_i} \times \left[S(t_i) \right]^{1-\delta_i} .$$

Usando las ecuaciones 3.3 y 3.4 tendremos:

$$L = \prod_{i=1}^{M} \left[\lambda_0(t_i) \exp(Z_i\beta) \right]^{\delta_i} \times \left[\exp(-\Lambda(t_i) \exp(Z_i\beta) \right]^{1-\delta_i} ,$$

introduciendo la variable

$$\mu_i = \Lambda(t_i) \exp(Z_i\beta) ,$$

obtenemos

$$L = \prod \left[\exp(\mu_i) \cdot \mu_i^{\delta_i} \right] \left[\lambda_0(t_i) / \Lambda(t_i) \right] .$$

Al primer factor de la función de verosimilitud, se le llama "núcleo" de la función de verosimilitud para δ_j variables distribuidas Poisson independientes con media μ_j.

El segundo factor solamente involucra (en general) parámetros desconocidos de forma, es decir, no involucra el vector de parámetros β. El modelo log-lineal para la funcion de riesgo implica un modelo log-lineal para la media de una variable Poisson definida como:

$$\ln \mu_i = \ln \Lambda(t_i) + Z_i\beta .$$

Si suponemos que tenemos α_j parámetros desconocidos ($j=1...s$); para dar una estimación de los parámetros y el vector β_j se utiliza el proceso iterativo de maximización de la función del Log de la verosimilitud, este se realiza utilizando el paquete GLIM.

Para esto se dan valores iniciales de los parámetros y se introducen en la función $\lambda(t)$ considerando esta función conocida. Introduciendo $\ln \Lambda(t_i)$ como un offset [6'], (que esta incorporado en el modelo log-lineal para el modelo de Poisson en GLIM), se obtiene el estimador de β_j. Para este estimador de β el estimador de m.v. de los α_j parametros desconocidos en μ_i pueden ser obtenidos de las ecuaciones de verosimilitud para esos parámetros, y así sucesivamente hasta la convergencia.

Para una función paramétrica de riesgo general este proceso es bastante complicado, y por lo tanto solamente la aplicaremos para algunos casos particulares.[7].

3.1.1 Distribucion Exponencial

Supondremos que $t>0$ y que

$$\lambda(t)=\exp(Z\beta)\cdot\lambda_0(t),$$

$\lambda_0(t)$=constante

donde

$$\eta = Z\cdot\beta ,$$

para este caso se tendrá $\Lambda(t)=t$;

de aquí se obtiene

$$f(t)=\exp\{Z\beta - t\exp(Z\beta)\} \qquad t > 0 ,$$

$$f(t)=0 \qquad t < 0 ,$$

por tanto

$$\ln \mu_i = \ln t_i + Z_i\beta ;$$

dado que $\frac{\lambda(t)}{\Lambda(t)}=\frac{1}{t}$ no tiene parámetro desconocido, la maximización de la función de verosimilitud, se reduce a ajustar un modelo de Poisson con media:

$$\mu_i = t_i\exp(Z\beta)$$

y

$$\eta = Z\cdot\beta .$$

3.1.2 Distribución Weibull

Para la distribución Weibull, suponemos $\Lambda(t)=t^{\alpha}$, $\alpha>0$ y obtenemos una función de riesgo proporcional a $\alpha t^{\alpha-1}$ y una función de densidad Weibull asociada; dada por:

$$f(t)= \alpha t^{\alpha}\left[\exp(Z\beta - t^{\alpha}\cdot\exp(Z\beta)\right] \qquad t > 0 ,$$

$$f(t)= 0 \qquad t \leq 0 ,$$

Además $\frac{\lambda(t)}{\Lambda(t)}=\frac{\alpha}{t}$ depende del parámetro desconocido α el cual debe ser estimado conjuntamente con las β . Así mismo se tiene $\mu_i=t_i^{\alpha}\exp(Z_i\beta)$). El log de la función de verosimilitud es:

$$\ln L = \ln\alpha + \sum_{i=1}^{M}(\delta_i \ln\mu_i - \mu_i) + \sum_{i=1}^{M}\delta_i \ln t_i \ ,$$

dado que el último término no depende de los parámetros puede omitirse.

Las ecuaciones de verosimilitud para β_j son:

$$\frac{\partial \Xi}{\partial \beta_j}=\sum_{i=1}^{M}(\delta_i - \mu_i) Z_{ij} =0$$

y la ecuación para α_j dado el vector β_j es:[3]

$$\frac{\partial \Xi}{\partial \alpha_j}=-\frac{r}{\alpha_j} + \sum_{i=1}^{M}(\delta_i - \mu_i) \ln t_i = 0$$

y por lo tanto

$$\hat{\alpha} =\left[\sum(\hat{\mu}_i - \delta_i) \ln t_i/r\right]^{-1} .$$

El procedimiento iterativo que se sigue para el cálculo de α comienza fijando $\alpha(0)=1$,es decir, se ajusta un modelo exponencial.El modelo Poisson es ajustado con un OFFSET que es $\alpha(0)=\ln(t)$ y los valores ajustados μ_{i0} son usados para reestimar dado un cierto $\alpha(0)'$. Un nuevo estimador $\alpha_1=\frac{\alpha_0+\alpha_0'}{2}$ se usa para definir un nuevo OFFSET $\alpha_1\log(t)$ para el modelo de Poisson y el proceso se continúa hasta la convergencia (ver apendice A).

3.1.3 Distribución valor extremo

Para esta distribución se tiene $\Lambda(t)=\exp(\alpha t)$, a ella le corresponde una función de riesgo proporcional a $\alpha\exp(\alpha t)$ y una función de densidad de la forma:

$$f(t_i)=\alpha\exp(\alpha t_i)\exp\{Z\beta - \exp(\alpha t + Z\beta)\} .$$

Para la estimación de los parámetros, se puede notar que haciendo la transformación $v=\exp(t)$, la distribución anterior se convierte en la distribución Weibull para la cual se utiliza el mismo proceso iterativo excepto que t se reemplaza por log t.

3.1.4 Distribución logística

Para considerar el ajuste del modelo logístico y Log-Logístico en GLIM; para datos censurados, se supone que los tiempos observados siguen la distribución logística:

$$f(t_i) = \frac{\alpha \exp(\alpha t_i + z\beta)}{1 + \exp(\alpha t_i + z\beta)} ,$$

donde α es un parámetro que se supone es el mismo para todas las observaciones.

De las observaciones se tiene t_M de los cuales r son censurados y (M-r) no censurados con un mecanismo de censura independiente.

Usando la función de densidad para t_i se puede obtener una función de verosimilitud para α y β basada en los datos.

La relación entre μ_i y las variables concomitantes, se lleva a traves de la función liga y el predictor lineal dados por

$$\eta_i = \ln\left\{\frac{\mu_i}{M-\mu_i}\right\} , \qquad (3.5)$$

$$\eta_i = \alpha t_i + z\beta , \qquad (3.6)$$

este modelo puede ser ajustado en GLIM usando errores binomiales con una liga logito, y suponiendo que αt_i es un OFFSET fijo, el valor de α puede ser estimado iterativamente. Un estimador inicial α_0 es ajustado usando un modelo normal para los datos no censurados solamente y tomando la varianza normal igual a la varianza logística

$\frac{\pi}{3\alpha^2}$, usando la ecuación (3.6) con ese valor de α_0 se obtiene una estimación de β y nuevamente se reestima α resolviendo la ecuación con $\frac{\partial \Xi}{\partial \alpha} = 0$ así hasta la convergencia.[9]

3.1.5 Modelo Paramétrico general

Los modelos propuestos anteriormente, se pueden presentar de manera resumida de la siguiente forma:

Primeramente para las distribuciones Weibull y Valor Extremo, construiremos la función del logaritmo de verosimilitud denotado por Ξ considerando que :

I Para la distribución Weibull se tiene

$$a)\ \lambda(t,Z) = \alpha t^{\alpha-1} \exp(Z\beta)\ ,$$

$$b)\ f(t) = \alpha t^{\alpha-1} \exp\{-\exp(Z\beta)t^{\alpha} + Z\beta\},$$

$$c)\ S(t) = \exp\{-\exp(Z\beta)\cdot t^{\alpha}\}\ .$$

Si tomamos en cuenta que se tienen tiempos $t_1, \ldots t_M$, algunos de los cuales son censurados y, que la verosimilitud es:

$$L = \prod_i \left[f(t_i)\right]^{\delta_i}\left[S(t_i)\right]^{1-\delta_i}\ ; \quad \text{con } \delta_i = \begin{cases} 0 \text{ si } i \text{ es censurado}, \\ 1 \text{ si } i \text{ es falla}, \end{cases}$$

$$\Xi = \ln L = \sum_i \left\{\delta_i \ln f(t_i) + (1-\delta_i)\ln S(t_i)\right\},$$

$$\delta_i \ln f(t_i) = \delta_i\left\{\ln\alpha + (\alpha-1)\ln t_i + Z\beta - t_i^{\alpha}\exp(Z\beta)\right\} \qquad (3.7)$$

$$(1-\delta_i)\ln S(t_i) = -(1-\delta_i)\exp(Z\beta)t_i^{\alpha}\ , \qquad (3.8)$$

sumando 3.7 y 3.8 tenemos

$$\Xi = \sum\left\{\delta_i \ln\alpha + (\alpha-1)\ln t_i + \delta_i Z\beta - \exp(Z\beta)\cdot t_i^{\alpha}\right\},$$

$$= d\ln\alpha - \sum_i \delta_i \ln t_i + \alpha\sum_i \delta_i \ln t_i + \sum_i \delta_i Z\beta - \sum_i \mu_i$$

$$= d\ln\alpha - \sum_i \delta_i \ln t_i + \sum_i\left\{\delta_i \ln\mu_i - \mu_i\right\},$$

entonces

$$\mu_i = t_i^{\alpha}\exp(Z\beta)\ ,$$

$$\eta_i = \ln\mu_i$$

$$\eta_i = \ln t_i + Z\beta\ .$$

II Para la distribución Valor Extremo se tiene

a) $\lambda(t,Z) = \alpha \exp\{\alpha t + Z\beta\}$,

b) $f(t) = \alpha \exp\{\alpha t + Z\beta - \exp(\alpha t + Z\beta)\}$,

c) $S(t) = \exp\{-\exp(\alpha t + Z\beta)\}$,

por lo tanto,

$$\Xi_0 = \ln L = \sum \{\delta_i \ln f(t_i) + (1-\delta_i) \ln S(t_i)\} .$$

Además

$$\delta_i \ln f(t_i) = \delta_i [\ln \alpha + \alpha t_i + (Z\beta - \exp(\alpha t_i + Z\beta)] \qquad (3.9)'$$

$$(1-\delta_i) \ln S(t_i) = (1-\delta_i)(-\exp(Z\beta + \alpha t_i) , \qquad (3.10)$$

sumando 3.9 y 3.10 queda:

$$\Xi_0 = d \cdot \ln \alpha + \Sigma [\delta_i \ln \mu_i - \mu_i] ,$$

con $\mu_i = \exp(\alpha t_i + Z\beta)$

y $\eta_i = \ln \mu_i \quad ; \quad d = \Sigma \delta_i ,$

el predictor lineal será:

$$\eta_i = \alpha t_i + Z\beta .$$

Puede notarse que Ξ_1 y Ξ_0 difieren en un término que es $-\Sigma \delta_i \ln t_i$ que no depende de los parámetros desconocidos y puede omitirse en la maximización de esta manera pueden considerarse los dos modelos en en análisis conjunto como sigue:

Para Weibull y Valor Extremo se tiene:

$$\Xi = \sum_i (\delta_i \ln \mu_i - \mu_i) + d \ln \alpha ,$$

i) donde δ_i tiene el valor 0 para censura y 1 para valores observados

ii) d, número total de valores observados

iii) la función liga es $\eta_i = \ln \mu_i$

iv) el predictor lineal es $\eta_i = U_i \alpha + Z \beta$ donde son los parámetros desconocidos y Z_{i0} es el vector unitario (1,0,0,..0) (K-entradas.)

v) $U_i = \begin{cases} t_i \text{ para valor extremo,} \\ \ln t_i \text{ para Weibull.} \end{cases}$

Un procedimiento análogo, se sigue para las distribuciones logística y log-logística quedando el logaritmo de la verosimilitud como: (Roger y Peacock 1980)

$$\Xi = \sum_i \left[\delta_i \ln \mu_i + (\omega_i - \delta_i) \ln (\omega_i - \delta_i) \right] + d \ln \alpha ,$$

donde

i) δ_i tiene el valor 0 para censura por la izquierda y 1 para valores observados y censuras por la derecha
ii) ω_i tiene el valor 1 para censuras por la derecha y por la izquierda y el valor 2 para valores observados
iii) d es el número total de valores observados
iv) la función liga logito:

$$\eta_i = \ln \left(\frac{\mu_i}{\omega_i - \mu_i} \right) ,$$

v) el predictor lineal es

$$\eta_i = U_i \alpha + \Sigma_i z \beta ,$$

donde α y β son parámetros desconocidos y z_{io} es el vector unitario $(1, 0, \ldots, 0)$.

v) U_i toma el valor de t_i para la logística, y $U_i = \log t_i$ para la log-logística.

Ambos modelos se ajustan en GLIM tomando en cuenta que el primer término de la verosimilitud involucra μ_i y se reduce a un ajuste con error Poisson y función liga log para la función Weibull y Valor Extremo, el otro modelo a un error binomial y una función liga logito para la función logística y log-logística.

El término extra $d \log \alpha$ se incluye como un offset para cada tiempo, el parámetro de escala es la unidad y el parámetro α se ajusta por un proceso iterativo usando varias veces la instruccion FIT.

Los modelos propuestos anteriormente pueden ser englobados en un modelo general; el procedimiento utilizado para su implementacion en GLIM evita un proceso iterativo de ajuste de los parámetros, así son ajustados α y β_j simultáneamente.

El modelo general que involucra las funciones logística, valor extremo, log-logística y Weibull se propone con:

$$\Xi = \sum_{i=1}^{M} L_i = \sum_{i}^{M} f_i (\mu_i) + d \cdot \ln \alpha ,$$

en la cual la $f_i(\mu_i)$ es una función conocida de μ_i y el predictor lineal; $\eta_i = U_i\alpha + \sum Z_{ij}\beta_j$
su implementación en GLIM puede llevarse a cabo mediante el siguiente procedimiento.

a) colocar el número de unidades como el número de observaciones mas uno, y leer los datos adecuadamente.
b) definir el vector "C" asignando los siguientes valores:

$$C = \begin{cases} 0 & \text{para las censura por la derecha,} \\ 1 & \text{observaciones ,} \\ 2 & \text{para censura por la izquierda,} \end{cases}$$

c) poner el vector

$$U \begin{cases} t \text{ para la logística, valor extremo} \\ \log(t) \text{ para log-logística, Weibull ,} \end{cases}$$

d) poner el último elemento de todas las variables ajustadas que tengan el valor de cero (%NU).

Existen dos subrutinas llamadas macros SETW y SETL: la primera para la distribución Weibull y Valor Extremo; y la segunda para la logística y log-logística, las cuales ajustan el modelo nulo(sin covariables). Posteriormente con la instruccion FIT se ajustan las covariables. (ver apéndice B)

Con la instrucción DISPLAY se muestra la estimación de los parámetros, residuales, etc.

2 MODELO SEMI-PARAMETRICO

Para considerar inferencias acerca de β, cuando la función $\lambda_0(t)$ es completamente desconocida, se tiene el modelo de riesgos proporcionales, este modelo fue propuesto por Cox (1972), en los siguientes casos.

3.2.1 Datos no censurados

Sea $\tau_1 < \tau_2, \ldots, < \tau_M$ que representan los tiempos de falla ordenados de los n individuos en muestra y, denótese por I_j al sujeto cuyo tiempo de falla es t_j. Así, I_j será igual a i si y sólo si $\tau_j = t_j$. Se define el conjunto $R(\tau_j) = \{i;\ t_i > t_j\}$, como el conjunto de individuos en riesgo en el instante previo a la ocurrencia del j-ésimo tiempo de falla ordenado y r_j se utiliza para representar su tamaño.

Para la derivación de la verosimilitud se considera que los conjuntos $\{\tau_j\}$ y $\{I_j\}$ son conjuntamente equivalentes a los datos originales o sea a los tiempos desordenados t_i. En ausencia del conocimiento de $\lambda_0(t)$ el conocimiento de los t_i pueden proporcionar poca o ninguna información acerca de dado que su distribución depende ampliamente de $\lambda_0(t)$. Así, debemos enfocar nuestra atención sobre los I_j; la distribución conjunta $p(i_1, i_2, \ldots, i_M)$ sobre todas las posibles permutaciones de $(1,2,\ldots,n)$; pueden ser derivadas explícitamente en este caso.

La probabilidad condicional de que $I_j = i$ dada la historia completa

$$h_j = \{\tau_1, \tau_2, \ldots, \tau_j; i_1, i_2, \ldots, i_{j-1}\},$$

posterior al j-esimo tiempo ordenado τ_j puede calcularse como la probabilidad condicional de que i falle al tiempo τ_j dado que un individuo del conjunto de riesgo $\mathcal{R}(\tau_j)$ falla en τ_j, es:

$$p(I_j = i \mid h_j) = \frac{\exp(z_i \beta)}{\sum_{k \in \mathcal{R}(\tau_j)} \lambda_k(\tau_j)},$$

la función $\lambda_0(\tau)$ se cancela debido a la suposición del modelo multiplicativo $[\lambda(\tau) = \lambda_0(\tau)\exp(Z\beta)]$, la función anterior es independiente de los $\tau_1, \ldots \tau_j$ es decir es igual a $p(i \mid i_1, i_2, \ldots, i_{j-1})$ que es la probabilidad de I_j dadas unicamente $I_1 = i_1$, $I_2 = i_2, \ldots, I_{j-1} = i_{j-1}$

Así, la distribución conjunta $p(i_1, \ldots, i_M)$ puede obtenerse como.

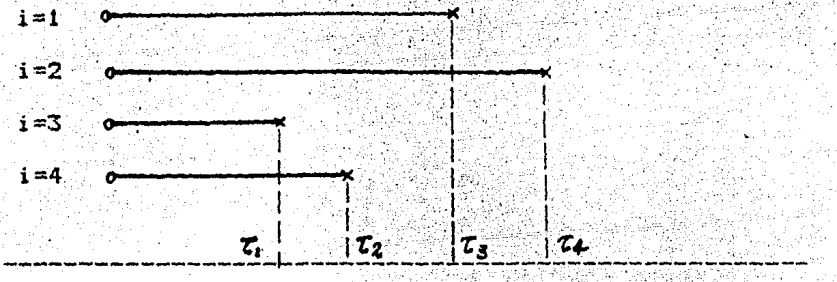
$$p(i_1,\ldots,i_M) = \prod_{j=1}^{M} p\,(i_j \mid i_1,\ldots,i_{j-1}) ,$$

$$= \prod_{j=1}^{M} \left[\frac{\exp(Z_{ij}\beta)}{\sum_{k \in \mathcal{R}(\tau_j)} \exp(Z_k \beta)} \right] . \qquad 3.11$$

como ejemplo considere la siguiente situación, mostrada en la figura:

Figura 3.1

Falla para cuatro individuos sin censura
Tiempos de falla $t_1, \ldots, t_4$ y niveles de cada individuo.



Los conjuntos de riesgo son:

$$\mathcal{R}(\tau_1)=\{1,2,3,4\}\,;\ \mathcal{R}(\tau_2)=\{1,2,4\}\,;\ \mathcal{R}(\tau_3)=\{1,2\}\,;\ \mathcal{R}(\tau_4)=\{2\}$$

En este caso $I_1=3$ $I_2=4$ $I_3=1$ $I_4=2$ así la probabilidad es:

$$p(3,4,1,2) = \frac{\exp(Z_3\beta)}{\sum_{k \in \mathcal{R}(\tau_1)} \exp(Z_k\beta)} \times \frac{\exp(Z_4\beta)}{\sum_{k \in \mathcal{R}(\tau_2)} \exp(Z_k\beta)} \times \frac{\exp(Z_1\beta)}{\sum_{k \in \mathcal{R}(\tau_3)} \exp(Z_k\beta)} \times \frac{\exp(Z_2\beta)}{\sum_{k \in \mathcal{R}(\tau_4)} \exp(Z_k\beta)} .$$

3.2.2 Datos censurados

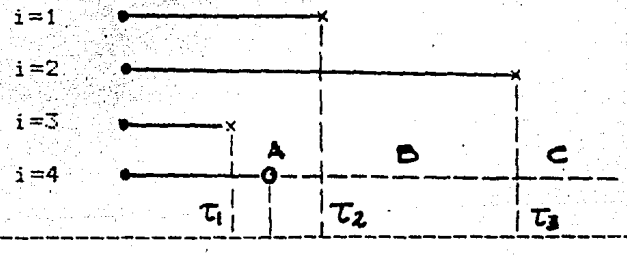
Supongase que existen d fallas observadas de una muestra de tamaño n, y sean $t_1 < t_2 <,\ldots,<t_n$ los tiempos de falla ordenados; nuevamente sea $I_j=i$ si el sujeto i falla al tiempo t_j ,y sea $\mathcal{R}(t_j)=\{i, t_i > t_j\}$ el conjunto de riesgo correspondiente de tamano r_j . En este caso la verosimilitud [10] será:

$$L = \prod_{j=1}^{r} \left[\frac{\exp(Z_{j_i}\cdot\beta)}{\sum_{k\in\mathcal{R}(t_j)} \exp(Z_k\beta)} \right] . \qquad 3.12$$

La ecuación 3.12 puede ser obtenida como la suma de todas las probabilidades 3.11 consistentes con el patrón observado de fallas y censuras. Como ejemplo considerese el siguiente diagrama:

Figura 3.2

falla de cuatro individuos con censura
x;falla,O;censura, t_1, t_2, t_3,tiempos de falla;



los conjuntos de riesgo son: $R_1=\{1,2,3,4\}$ $R_2=\{1,2\}$ $R_3=\{2\}$ A,B y C posibles posiciones para el tiempo de falla del individuo censurado.

Para este caso la verosimilitud será:

$$L = \frac{\exp(Z_3\beta)}{\sum_{k\in\mathcal{R}(t_3)} \exp(Z_k\beta)} \times \frac{\exp(Z_1\beta)}{\sum_{k\in\mathcal{R}(t_2)} \exp(Z_k\beta)} \times \frac{\exp(Z_2\beta)}{\sum_{k\in\mathcal{R}(t_1)} \exp(Z_k\beta)} , \quad [11]$$

la cual es la suma de los tres términos correspondientes posiciones posibles para τ_4 relativas a τ_1, τ_2, τ_3, a saber

A $$\frac{\lambda(3)}{K_1(K)} X \frac{\lambda(4)}{K_2(K)} X \frac{\lambda(1)}{K_3(K)} X \frac{\lambda(2)}{K_4(K)} ,$$

B $$\frac{\lambda(3)}{K_1(K)} X \frac{\lambda(1)}{K_2(K)} X \frac{\lambda(4)}{K_3(K)} X \frac{\lambda(2)}{K_4(K)} ,$$

C $$\frac{\lambda(3)}{K_1(K)} X \frac{\lambda(1)}{K_2(K)} X \frac{\lambda(2)}{K_3(K)} X \frac{\lambda(4)}{K_4(K)} ,$$

3.2.3 Fallas multiples

Las verosimilitudes halladas anteriormente, no son estrictamente apropiadas para tiempos de supervivencia discretas, las cuales pueden involucrar multiplicidades; en este caso pueden usarse dos aproximaciones. Primero si la escala es realmente discreta el modelo de riesgos proporcionales puede ser reemplazado por un modelo logístico discreto
donde $\lambda(t,Z)$ representa ahora:

$$Pr(T < t+1 \mid T > t),$$

para un individuo con variable explicativa Z.
Para derivar la verosimilitud para β , sean $\tau_1 <,\ldots,<\tau_r$ los r diferentes tiempos de falla ordenados y sea g la multiplicidad de fallas en el tiempo τ_K . La historia h ahora incluye las multiplicidades para todos los tiempos de falla mayores incluyendo τ_j . La probabilidad condicional de que $i_1, i_2, \ldots i_r$ fallen del conjunto de riesgo $R(\tau_j)$ dado h_j es:

$$P\ (i_1, \ldots, i_g | h_j) = \frac{\lambda(i_1)\cdot\lambda(i_2)\cdots\lambda(i_g)}{\sum_{K \in S(j,g)} \lambda(K_1)\cdots\lambda(K_g)} , \qquad 3.13$$

donde S(j,g) denota el conjunto de todas las selecciones de $g=g_j$ entradas del conjunto de riesgo $\mathcal{R}(t_j)$ de tamaño $r_j=r$,esta 3.13 es la contribución para un sólo tiempo de falla [12].
La verosimilitud queda

$$L = \prod_{j=1}^{g} \left[\frac{exp(S_j \cdot \beta)}{\sum_{K \in S(i,g)} exp(S_{jK} \cdot \beta)} \right] ,$$

donde $S_j = Z_{j1} + Z_{j2} + \ldots + Z_{jg}$ es la suma de las vectores Z sobre los individuos que realmente fallan a τ_j , cada S_{jK} es la correspondiente suma sobre una g-ada $(K_1, \ldots, K_g)$ de sujetos que pueden haber fallado al tiempo τ_j .

La segunda aproximaciones consiste en sumar todos los términos de la verosimilitud marginal 3.12 del modelo contínuo que son consistentes con los datos observados. Por ejemplo si i=1,2 son fallas al tiempo t de i=1,2,3,4 en riesgo la contribución de t a esta verosimilitud sera:

$$\frac{\lambda(1)}{K_1(k)} X \frac{\lambda(2)}{K_2(k)} + \frac{\lambda(2)}{K_1(k)} X \frac{\lambda(1)}{K_2(k)} .$$

Estas expresiones tienen el problema de ser muy difíciles de calcular. Si g es muy grande comparado con r.

Una buena aproximación es considerar en la verosimilitud todas las sumas en el denominador para incluir todas las entradas en el correspondiente conjunto de riesgo. En el ejemplo anterior quedara:

$$\frac{2\,\lambda(1)\,\lambda(2)}{[K_1(k)]^2} .$$

En general se tendrá: (Peto. discusion articulo de Cox 1972):

$$\frac{g! \cdot \lambda(i_1) \cdots \lambda(i_g)}{\left[K_j(k)\right]^g} ,$$

en la cual es mas apropiada desde el punto de vista de cálculo y es bastante satisfactoria excepto cuando existen demasiadas fallas múltiples comparadas con el conjunto de riesgo.[13]

Si $g_j = 1$ se recupera la verosimilitud 3.12 .

3.2.4 Implementación en GLIM

Para la implementación del modelo de Cox (1972) en GLIM, usaremos la formulación de Whitehead (1980), primeramente para el caso de fallas y censuras distintas y posteriormente para fallas múltiples.

Consideramos datos que consisten en tiempos de supervivencia de los cuales algunos son censurados y otros corresponden a fallas. Supongase que existen r muertes que ocurren después de los tiempos $t_1,\ldots,t_r$ y supóngase que todos esos tiempos son distintos. Consideremos la muerte que ocurre despues del tiempo t_h y sea N_h el número de tiempos de supervivencia, censurados o no censurados los cuales son mayores de t_h. Además consideremos que esos N_h pacientes caen en k grupos; el grupo j-ésimo consiste de N_{hj} pacientes todos con vector de variables explicativas $z_{hj}=z_{hj}(t)$ $(j=1,\ldots k(h))$ y además $N_{h1}+N_{h2}+\ldots+N_{hk}=N_h$ $(h=1,\ldots,r.)$

Supóngase que la muerte al tiempo t_h la cual sucede a uno de esos N_{hj*} pacientes, esta involucrada en el j*-ésimo grupo.

Como ya hemos visto, el modelo de riesgos proporcionales aborda el problema de la estimación de los parámetros asociados, en ausencia del conocimiento de $\lambda_j(t)$, mediante la verosimilitud parcial, dada por:

$$L(z;\beta) = \prod_{h=1}^{r} \frac{\exp(z_{hj*}\cdot\beta)}{\sum N_{hj}\exp(z_{hj}\beta)}. \qquad 3.15$$

Esta expresión puede ser usada para su implementación en GLIM considerando un modelo de Poisson apropiado. La idea es considerar una verosimilitud que sea proporcional a la expresión dada por Cox. El modelo puede ser ajustado en GLIM usando el error Poisson y asi obtener el estimador máximo verosímil de β.

El modelo Poisson puede ser descrito como sigue: para cada valor de h, de 1 a r, sea $X_{h1},\ldots,X_{hK}$ variables aleatorias asociadasde Poisson independientes, donde X_{hj} tiene parámetro μ_{hj} con :

$$\mu_{hj}=N_{hj}\exp(\alpha_h+z_{hj}\cdot\beta) \quad , \quad (j=1,\ldots,k), \quad (3.16)$$

α_h son constantes que determinan el comportamiento de la función de supervivencia. Si $X_{hj=K(h)}$ toma el valor de 1 y los otros el valor de cero, entonces la verosimilitud de α_h y β basada en $X_{h1},\ldots,X_{hK(h)}$ es

$$\mu_{hj}\exp\left(-\sum_j \mu_{hj}\right)=\frac{N_{hj}\exp(\alpha_h+Z_{hj}\cdot\beta)}{\exp\left\{\sum N_{hj}\exp(\alpha_h+Z_{hj}\beta)\right\}}, \qquad 3.16$$

donde además se debe cumplir $\sum \hat{\mu}_{hj}=1$, tal que

$$\exp(\hat{\alpha}_h)=\left\{\sum_j N_{hj}\exp(Z_{hj}\cdot\beta)\right\}^{-1},$$

sustituyendo en (3.16) obtenemos

$$\prod_{h=1}^{r}\mu_{hj}\exp\left(-\sum\mu_{hj}\right)=\prod_{h=1}^{r}\left[\frac{N_{hj}\times\exp(Z_{hj}\cdot\beta)\times\left\{\sum N_{hj}\exp(Z_{hj}\hat{\beta})\right\}^{-1}}{\exp\left\{\exp(\hat{\alpha}_h)\sum_j N_{hj}\exp(Z_{hj}\cdot\hat{\beta})\right\}}\right],$$

usando (3.16)

$$L(z,\beta)=\prod_{h=1}^{r}\left\{\frac{N_{hj}\exp(Z_{hj}\cdot\hat{\beta})\left[N_{hj}\exp(Z_{hj}\cdot\beta)\right]^{-1}}{\exp(1)}\right\}$$

$$L(z,\beta)=\prod_{h=1}^{r}\left\{\frac{N_{hj}\cdot\vec{e}^{1}\cdot\exp(Z_{hj}\cdot\beta)}{\sum N_{hj}\exp(Z_{hj}\cdot\beta)}\right\}.$$

Esta es la verosimilitud para α y β basados en X_{hj} (j=1,...k, h=1,...,r) que es proporcional al máximo obtenido de la verosimilitud de la ecuación (3.14). Por tanto, el modelo Poisson y el modelo de supervivencia tendrá estimadores idénticos para β e iguales cocientes de verosimilitud.

Fallas múltiples

Para el análisis de fallas múltiples consideremos lo siguiente: sean nuevamente $t_1,t_2,\ldots,t_r$ tiempos de supervivencia distintos para los cuales puede suceder que existan fallas múltiples $(m_h>1)$ después de cada t_h (h=1,...,r) en donde $m_1+\ldots+m_r$ es el total de fallas.

Para la implantación en GLIM, en este caso, utilizaremos la aproximación de Cox y Peto ; como a continuación se muestra.

3.2.4.1 Aproximación de Peto

Consideraremos que se tienen j grupos (j=1,...,k) y que m_h de las muertes que ocurren al tiempo t caen en el j-ésimo grupo, cuyas variables explicativas son:

$$Z_{hj} \quad \text{con } j=1,2,\dots,K(h)\ ;\ h=1,2,\dots,r;\ m_{h1}+\cdots+m_{hK}=m_h,$$

en estas condiciones la verosimilitud segun la generalizacion de Peto es:

$$L=\prod_{h=1}^{r}\frac{\exp\left(\sum m_{hj}\cdot Z_{hj}\beta\right)}{\binom{N_h}{m_h}\left\{\sum N_{hj}\exp\left(Z_{hj}\cdot\beta\right)/N\right\}^{m_h}}, \qquad (3.17)$$

para asociarle un modelo de Poisson, supongamos $X_{hj}=m_{hj}$ (j=1,...,k; h=1,...,r) y la verosimilitud de α_h y β será:

$$=\prod_{h=1}^{r}\frac{\exp\left(\sum m_{hj}\,Z_{hj}\cdot\beta\right)}{\binom{N_h}{m_h}\left\{\sum N_{hj}\cdot\exp\left(Z_{hj}\cdot\beta\right)/N_h\right\}^{m_h}},$$

además en su máximo la verosimilitud cumple:

$$\sum_j N_{hj}\exp\left(\hat{\alpha}_h+\ Z_{hj}\cdot\hat{\beta}\right)=m_h\ ,$$

entonces se tendrá

$$L=\prod_{h=1}^{r}\frac{\left\{\prod N_{hj}^{m_{hj}}\right\}e^{-m_h}\cdot\ \exp\left(\sum m_{hj}\cdot Z_{hj}\cdot\hat{\beta}\right)}{\left\{\sum N_{hj}\exp\left(Z_{hj}\cdot\hat{\beta}\right)/m\right\}^{m_h}},$$

que es proporcional a (3.17). Así la aproximacion de Peto y su modelo asociado Poisson darán los mismos resultados para $\hat{\alpha}_h$ y β y los mismos cocientes de verosimilitudes. (ver apendice C.1)

3.2.4.2 Aproximación de Cox

Consideraremos las combinaciones de m_h individuos escogidos de N_h, con tiempos de supervivencia $> t_{hj}$, $\binom{N_h}{m_h} = M_h$ y definamos el vector asociado de variables explicativas de tal combinacion S como la suma de los valores correspondientes a los individuos en cuestión.

Supongase que M_{hj}, de todas las M_h, forman un grupo con vector comun de variables explicativas.

$$S_{hl}, \quad l = 1,\ldots,k\,;\; h = 1,\ldots,r\,;\; M_{h1}+\cdots+M_{hk} = M_h .$$

Consideremos las combinaciones correspondientes a las m_h muertes que estan en el grupo $l^*(h)$ cuyo vector de covariables es $S_{hl^*(h)}$. Según Cox se puede obtener $\hat{\beta}$ maximizando la verosimilitud parcial dada por:

$$L = \prod_{h=1}^{r} \frac{\exp(S_{hl}\cdot\beta)}{\sum_{l} M_{hl}\exp(S_{hl}\cdot\beta)} .$$

El modelo asociado Poisson tiene $X_{h1},\ldots,X_{hk}$ como variables aleatorias independientes $(h=1,\ldots,r)$ y X_{hl} tiene asociada:

$$\mu_{hl} = M_{hl}\cdot\exp(\alpha_h - S_{hl}\beta) .$$

Si X_{hl} vale 1 y los otros valen cero, entonces la verosimilitud de α_h y β será:

$$\mu_{hl}\exp(-\textstyle\sum_l \mu_{hl}) = \frac{M_{hl}\cdot\exp(\alpha_h\; S_{hl}\beta}{\exp\left\{\sum_l M_{hl}\cdot\exp(\alpha_h + S_{hl}\beta)\right\}},$$

tomando en cuenta que:

$$\sum_l M_{hl}\cdot\exp(\hat{\alpha}_h + S_{hl}\hat{\beta}) = 1$$

donde $\exp(\hat{\alpha}_h) = \left[\sum_l M_{hl}\cdot\exp(S_{hl}\cdot\hat{\beta})\right]^{-1}$, obtenemos

$$L = \prod_{h=1}^{r} \frac{M_{hl}\exp(\hat{\alpha}_h + S_{hl}\cdot\hat{\beta})}{\exp\left[\sum M_{hl}\cdot\exp(\hat{\alpha}_h + S_{hl}\hat{\beta})\right]} .$$

$$L = \prod_{h=1}^{r} \left[\frac{\bar{e}^{1} \cdot M_{h\ell} \cdot \exp(S_{n\ell} \cdot \beta)}{\sum_{\ell} M_{h\ell} \cdot \exp(S_{n\ell} \cdot \beta)} \right].$$

que es proporcional a la aproximacion de Cox y da los mismos estimadores para β (ver apendice C.2).

CAPITULO IV

CALCULO DE LA FUNCION DE SUPERVIVENCIA

Para la estimación de la función de supervivencia de los casos analizados anteriormente,haremos uso de todos los resultados obtenidos en los capítulos anteriores. Primeramente analizaremos el caso paramétrico para las funciones particulares estudiadas, es decir las distribuciones Exponencial, Weibull y Valor extremo. Por último, analizaremos el caso semi-paramétrico propuesto por Cox.

Los datos que se emplean son los usados por Gehan(1965), para analizar los tiempos de remisión de pacientes con leucemia. Estos datos se muestran en el cuadro 4.1.

Cuadro 4.1

Tiempos de remisión (semanas) de pacientes con leucemia

Muestra 0 (tratamiento)	6*,6,6,6,7,9*,10*,10,11*,13,16,17* 19*,20*,22,23,25*,32*,32*,34*,35*
Muestra 1 (control)	1,1,2,2,3,4,4,5,5,8,8,8,8, 11,11,12,12,15,17,22,23

* censura

Usaremos inicialmente el estimador de Kaplan-Meier (ver referencia 12) para la función de supervivencia. Este estimador se utilizará como referencia para la comparación con los diversos modelos ajustados.

El estimador Kaplan-Meier se obtiene con la expresión:

$$\hat{S}_K(t)=\prod_{t_i \le t}\left(1-\frac{m_i}{r_i}\right),$$

donde m(i) es el número de fallas múltiples en el tiempo t(i) y r(i) es el número de individuos en el conjunto de riesgo $\mathcal{R}(t_i)$ para (t_i).

Por ejemplo el valor calculado para S (t) en el grupo control y t=3 es, tomando en cuenta que S(0)=1

$$\hat{S}_K(3) = \prod_{t_i<3}\left(1 - \frac{m_i}{r_i}\right) = \hat{S}_0(0)\left(1 - \frac{m_1}{r_1}\right)\left(1 - \frac{m_2}{r_2}\right)\left(1 - \frac{m_3}{r_3}\right) = .762$$

y así para todos los datos dados en cada grupo, se obtienen los valores de la función de supervivencia que se muestran en el cuadro 4.2 y en la su gráfica 4.1.

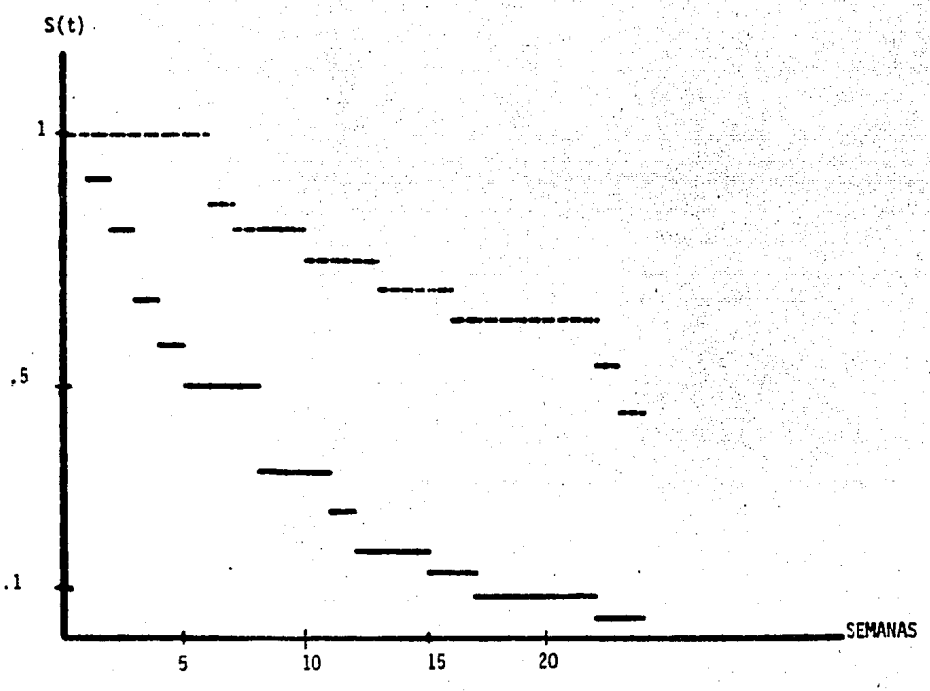
Cuadro 4.2

Estimación de la función de Supervivencia para el grupo tratado y el grupo control con el estimador KAPLAN-MEIER

INTERVALO	TIEMPOS DE FALLA MUESTRA (0)	TIEMPOS DE FALLA MUESTRA (1)	POBLACION EN RIESGO NUMERO EN (0)	POBLACION EN RIESGO NUMERO EN (1)	$S_j(t)$	$S_j(t)$
[0,1]		1,1	21	21	1	.90
[1,2]		2,2	21	19	1	.81
[2,3]		3	21	17	1	.76
[3,4]		4,4	21	16	1	.67
[4,5]		5,5	21	14	1	.57
[5,6]	6,6,6		21	12	.86	.57
[6,7]	7		17	12	.81	.57
[7,8]		8,8,8,8	16	12	.81	.38
[8,10]	10		15	8	.76	.38
[10,11]		11,11	13	8	.76	.29
[11,12]		12,12	12	6	.76	.19
[12,13]	13		12	4	.70	.19
[13,15]		15	11	4	.70	.14
[15,16]	16		11	3	.64	.14
[16,17]		17	10	3	.64	.09
[17,22]	22	22	7	2	.55	.05
[22,23]	23	23	6	1	.46	.04

Gráfica 4.1

Función de Supervivencia empírica para los datos del cuadro 4.1. Estimación del producto límite:---------, muestra 0 (6-mp):_____ muestra 1 (control).



4.1 Cálculo de la función de supervivencia en el caso Paramétrico.

La estimación de la función de supervivencia para el caso paramétrico se lleva a cabo directamente mediante la relación

$$S(t) = \exp\left[-\int \lambda_0(u)\,du\right]^{\exp(Z\beta)},$$

en la cual la función fundamental de supervivencia será:[14]

$$S_0(t) = \left[\exp(-\Lambda(t))\right]^{\exp(\beta_0)}$$

y

$$\Lambda(t) = \int_0^t \lambda_0(u)\,du\ ,$$

para los casos aquí considerados se tiene el siguiente cuadro:

Cuadro 4.3

Fórmula para la función de supervivencia para la muestra cero y muestra uno respectivamente de acuerdo a la función fundamental propuesta.

Función	$\Lambda(t)$	$S_0(t)$	$S(t)$
Exponencial	t	$\exp(-\exp(\beta_0)t)$	$\exp(-t\exp(Z\beta))$
Weibul	t	$\exp(-t^{\alpha}\exp(\beta_0))$	$\exp(-t^{\alpha}\exp(Z\beta))$
Valor Extremo	$\exp(t)$	$\exp(-\exp(\alpha t+\beta_0))$	$\exp(-\exp(\alpha t+Z\beta))$

Los parámetros calculados para cada caso se muestran en la siguiente cuadro junto con sus errores estándar.

Cuadro 4.4
Parámetros estimados de forma de la media general y del coeficiente de la variable tratamiento, asi como los errores estándar.

Función	α	e.e.	β_0	e.e	β_1	e.e.
Exponencial	1	-	2.92	.1979	1.53	.3958
Weibull	1.37	.1984	-3.934	.593	1.73	.4041
Valor Extremo	.1154	.01588	-2.396	.3897	2.172	.4534

En base a los datos obtenidos se calcularon los estimadores de las funciones de supervivencia que se muestran a continuación, al igual que sus gráficas respectivas.

Cuadro 4.5

Estimación de la función de Supervivencia para el grupo tratado y para el grupo control respectivamente, cuando la función fundamental es constante.

t	S(t) DEL GRUPO 0 *	S(t) DEL GRUPO 1**
1	.95	.78
2	.90	.61
3	.86	.47
4	.82	.37
5	.78	.29
6	.74	.22
7	.70	.17
8	.67	.14
9	.64	.11
10	.61	.08
11	.58	.06
12	.55	.0503
13	.52	.039
14	.50	.031
15	.47	.024
16	.45	.019
17	.43	.0145
18	.41	.0113
19	.39	.009
20	.37	.0069
21	.35	.0054
22	.33	.0042
23	.32	.0033

* Grupo 0 (tratamiento)
** Grupo 1 (control)

Gráfica 4.2

Estimación del producto Límite --------,
muestra 0 (6 mp),_____, muestra 1 (control)
Estimación obtenida por medio del modelo
exponencial: X, muestra 0, 0 muestra 1.

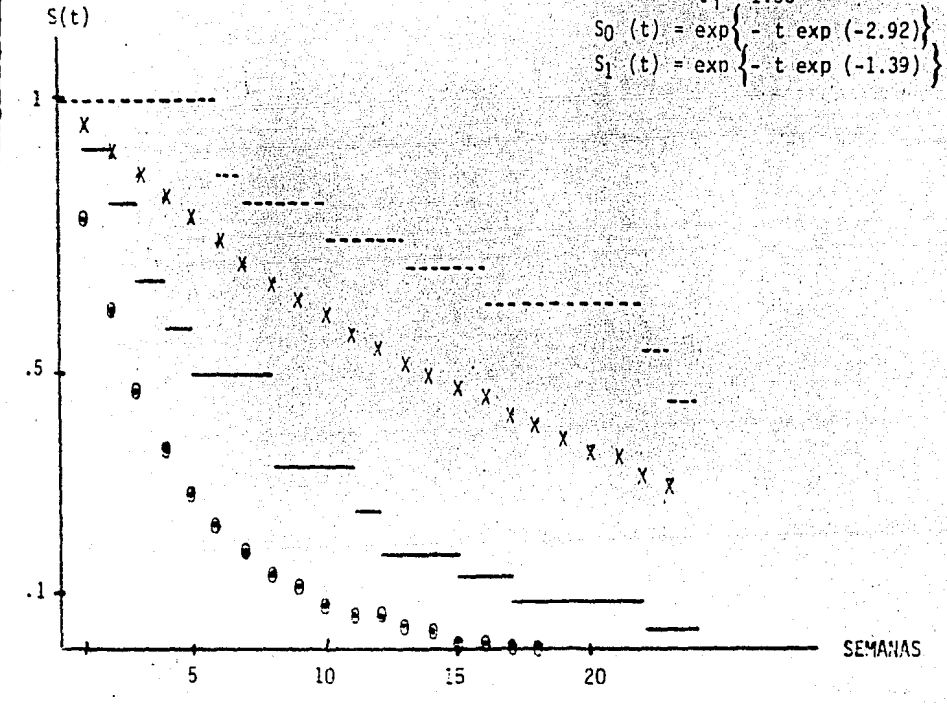
$$S(t) = \exp\left\{-t \exp(Z\ )\right\}$$

$$\beta_0 = -2.92$$

$$\beta_1 = 1.53$$

$$S_0(t) = \exp\left\{-t \exp(-2.92)\right\}$$

$$S_1(t) = \exp\left\{-t \exp(-1.39)\right\}$$



Cuadro 4.6

Estimación de la función de Supervivencia cuando la función de riesgo es Weibull;para el grupo al cual se le esta aplicando el tratamiento y al grupo control respectivamente.

t	S(t) DEL GRUPO 0 *	S(t) DEL GRUPO 1 **
1	.98	.90
2	.95	.75
3	.91	.61
4	.87	.48
5	.83	.37
6	.79	.28
7	.75	.21
8	.71	.15
9	.67	.11
10	.63	.08
11	.59	.05
12	.55	.04
13	.51	.02
14	.48	.017
15	.44	.01
16	.41	.0075
17	.38	.0049
18	.35	.0032
19	.32	.0020
20	.30	.0013
21	.27	.0008
22	.25	.0005
23	.23	.0003

* Grupo 0 (tratamiento)
** Grupo 1 (control

. Gráfica 4.3

Estimación del Producto Límite: -------,
muestra 0 (6-mp);_____, muestra 1 (control)
Estimación obtenida mediante el modelo Weibull:
X, muestra 0; 0 muestra 1.

$$S(t) = \exp\left\{-t^{\alpha}\exp(Z\beta)\right\}$$

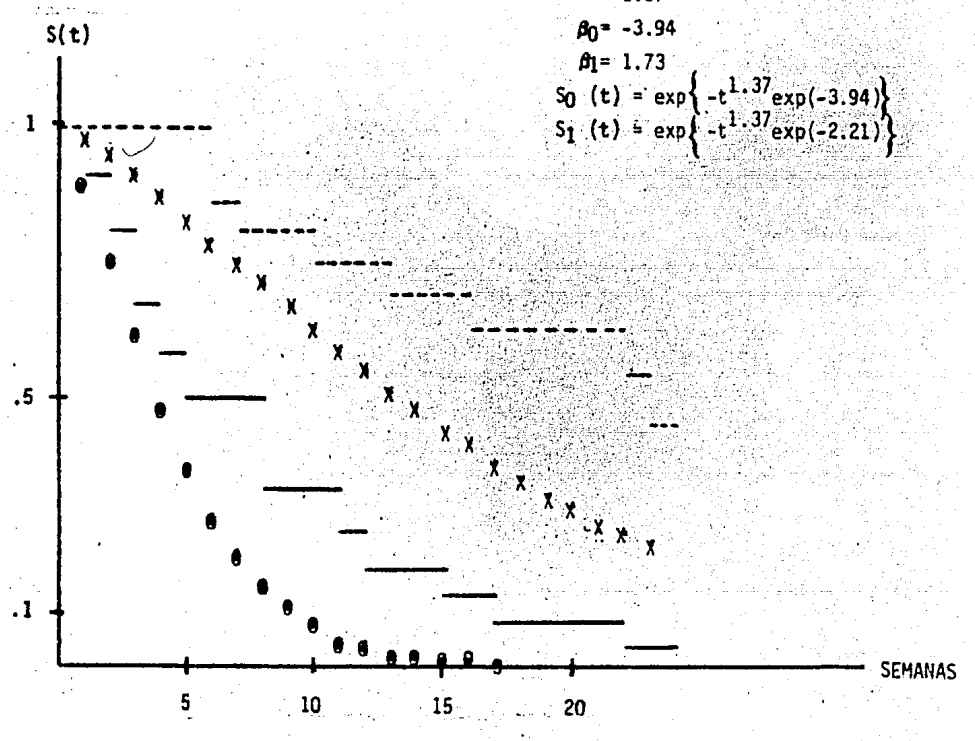
$\alpha = 1.37$

$\beta_0 = -3.94$

$\beta_1 = 1.73$

$$S_0(t) = \exp\left\{-t^{1.37}\exp(-3.94)\right\}$$

$$S_1(t) = \exp\left\{-t^{1.37}\exp(-2.21)\right\}$$



Cuadro 4.7

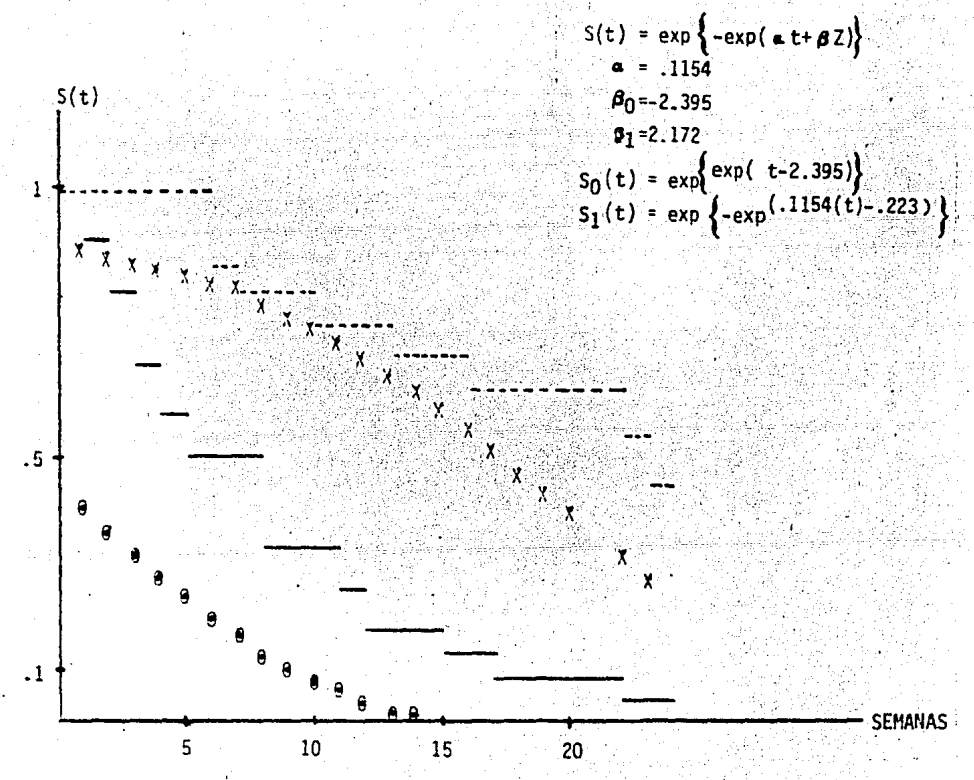
Estimación de la función de Supervivencia para la muestra tratada y la muestra control respectivamente; con una función fundamental valor extremo.

t	S(t) DEL GRUPO 0 *	S(t) DEL GRUPO 1 **
1	.90	.41
2	.89	.37
3	.87	.32
4	.86	.28
5	.85	.24
6	.83	.20
7	.83	.17
8	.79	.13
9	.77	.10
10	.75	.08
11	.72	.06
12	.69	.04
13	.66	.02
14	.63	.018
15	.60	.011
16	.56	.063
17	.52	.0034
18	.48	.0017
19	.44	.0008
20	.40	.0003
21	.36	.0001
22	.32	.00004
23	.27	.00001

* Grupo 0 (tratamiento)
** Grupo 1 (control)

Gráfica 4.4

Estimación del Producto límite:--------,
muestra 0 (6-mp); _____, muestra 1 (control)
Estimación obtenida por medio del modelo Valor
Extremo: X, muestra 0; 0 muestra 1.



4.2 Estimación de la función de supervivencia en el caso Semi-paramétrico

Para el cálculo de la función de supervivencia en el caso semi-paramétrico hay que considerar nuevamente que:

$$\hat{S}(t)=\exp\left\{-\int\lambda(t,Z)\,dt\right\}$$

o

$$\hat{S}(t)=\left[S_0(t)\right]^{\exp Z\hat{\beta}},$$

donde $\hat{S}_0(t)=\exp\left[-\int_0^t\lambda_0(u)\,du\right]$, para el caso analizado bajo las condiciones de que $\lambda_0(t)=cte$ por intervalo la función que estima $S_0(t)$ para el modelo de Peto es [15] :

$$\hat{S}_0(t)=\exp\left\{-\sum_{t_h\le t}\exp(\hat{\alpha}_h)\right\},$$

donde $\hat{\alpha}_h$ es el valor estimado de $\lambda(t_h)$ calculado con el paquete GLIM y que se muestra en el apéndice C.1.

Para hallar los valores de $S_0(t)$, por ejemplo para t=2

$$\hat{S}_0(2)=\exp\left\{\sum_{t_h\le 2}\exp(\hat{\alpha}_h)\right\},$$

$$\hat{S}_0(2)=\exp(-(\exp(\hat{\alpha}_1)+\exp(\hat{\alpha}_2))$$

$$\hat{S}_0(2)=.965$$

y así para todos los valores de $\hat{S}_0(t_h)$. Los valores obtenidos para $\hat{S}(t)$ en el caso de la aproximación de Peto se muestran en el cuadro 4.8 y en la gráfica 4.5.

Cuadro 4.8

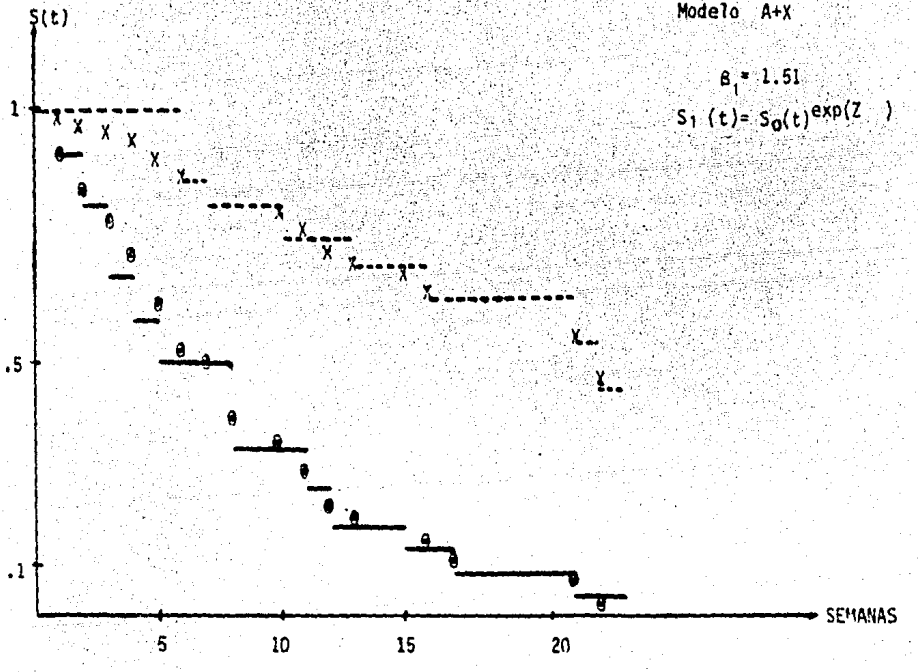
Estimación de la función de Supervivencia (con la aproximación de Peto) para el grupo que se está aplicando el tratamiento y para el grupo control.

INTERVALO	S(t) DEL GRUPO 0 *	S(t) DEL GRUPO 1 **
[0,1]	.98	.91
[1,2]	.96	.83
[2,3]	.95	.79
[3,4]	.93	.72
[4,5]	.90	.62
[5,6]	.87	.53
[6,7]	.86	.51
[7,8]	.81	.39
[8,10]	.79	.34
[10,11]	.76	.29
[11,12]	.72	.23
[12,13]	.70	.20
[13,15]	.68	.17
[15,16]	.65	.14
[16,17]	.62	.11
[17,22]	.55	.07
[22,23]	.45	.03

* Grupo 0 (tratamiento)
** Grupo 1 (control)

Gráfica 4.5

Estimación del producto Límite: -------,
muestra 0 (6-mp); _____, muestra 1 (control)
Estimación obtenida por la Aproximación de Peto
X, muestra 0; O muestra 1.



Para calcular $\hat{S}(t)$ en el caso de Cox se usa la fórmula

$$S(t)=\left[S_0(t)\right]^{\exp(Z_0\beta_0+Z_1\beta_1)},$$

la cual es análoga a la de Peto, salvo que en este caso la estimación para $S_0(t)$ se lleva a cabo mediante la fórmula:

$$\hat{S}_0(t)\cong\exp\left\{-\sum_{t_h\leq t}\frac{m_h}{N_h}\binom{N_h}{m_h}^{\frac{1}{m_h}}\exp\left(\frac{\hat{\alpha}_h}{m_h}\right)\right\} \quad [16],$$

en la cual:

$$m_h=m_{h1}+m_{h2}+\ldots+m_{hk},$$

$$N_h=N_{h1}+N_{h2}+\ldots+N_{hk},$$

$\hat{\alpha}_h$ –son los parámetros estimados mediante el modelo de Cox.

m_{hj} – denota el numero de muertes que suceden en el tiempo t, en el grupo j-ésimo ,

N_{hj} – es el número de individuos en el grupo j-ésimo, para el tiempo t.

Por ejemplo para t=3

$$\hat{S}_0(3)=\exp\left\{-\sum_{t_h\leq 3}\frac{m_h}{N_h}\binom{N_h}{m_h}^{\frac{1}{m_h}}\exp\left(\frac{\hat{\alpha}_h}{m_h}\right)\right\},$$

$$\hat{S}_0(3)=\exp\left\{-\frac{m_1}{N_1}\binom{N_1}{m_1}^{\frac{1}{m_1}}\exp\left(\frac{\hat{\alpha}_1}{m_1}\right)-\frac{m_2}{N_2}\binom{N_2}{m_2}^{\frac{1}{m_2}}\exp\left(\frac{\hat{\alpha}_2}{m_2}\right)-\frac{m_3}{N_3}\binom{N_3}{m_3}^{\frac{1}{m_3}}\exp\left(\frac{\hat{\alpha}_3}{m_3}\right)\right\},$$

$$\hat{S}_0(3)=\exp\left\{-\frac{2}{42}\binom{42}{2}^{\frac{1}{2}}\exp\left(-\frac{8.976}{2}\right)-\frac{2}{40}\binom{40}{2}^{\frac{1}{2}}\exp\left(\frac{-8.807}{2}\right)-\frac{1}{38}\binom{38}{1}^{\frac{1}{1}}\exp\left(\frac{-4.678}{1}\right)\right\}$$

$$\hat{S}_0(3)=\exp\ (-0.0157+.01708+.0093)$$

$$\hat{S}_0(3)=.9588\cong 96$$

y así sucesivamente para todos los valores de t considerados. Los resultados se muestran en el cuadro 4.9 y su gráfica correspondiente en la página 51 .

Cuadro 4.9

Estimación de la función de Supervivencia (con la aproximación de Cox) para el grupo tratado y para el grupo control.

INTERVALO	S(t) DEL GRUPO 0	S(t) DEL GRUPO 1
[0,1]	.98	.90
[1,2]	.97	.86
[2,3]	.96	.81
[3,4]	.94	.73
[4,5]	.92	.65
[5,6]	.89	.55
[6,7]	.88	.52
[7,8]	.83	.39
[8,10]	.82	.36
[10,11]	.79	.30
[11,12]	.75	.23
[12,13]	.73	.20
[13,15]	.70	.16
[15,16]	.68	.14
[16,17]	.65	.11
[17,22]	.58	.06
[22,23]	.47	.02

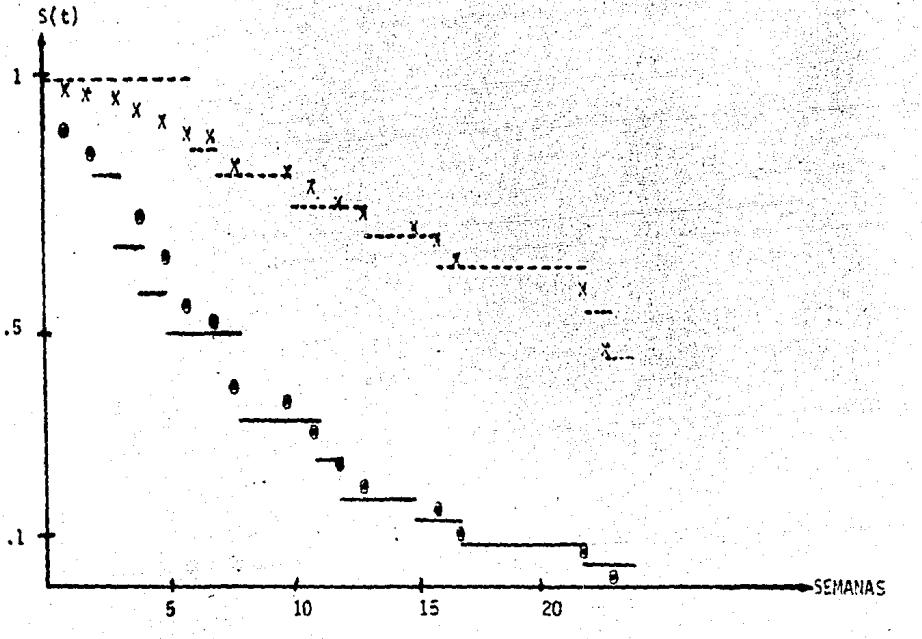
Gráfica 4.6

Estimación del producto límite:--------,
muestra 0 (á-mp); _____,muestra 1 (control).
Estimación con la aproximación de Cox:
X,muestra 0; 0 muestra 1.

Modelo A+X

β_1= 1.63

$S_1(t)=S_0(t)^{\exp Z\beta}$



4.3 Bondad de Ajuste

Existen varios enfoques (gráfico y estadístico) para analizar la bondad de ajuste de los datos a cierto modelo propuesto, tanto en el caso paramétrico como en el semi-paramétrico.

Para el análisis de la bondad de ajuste de los datos observados al modelo propuesto (que en el caso paramétrico es dar $\hat{S}$ (t) con todos los parámetros conocidos y en el caso semi-paramétrico consiste en hallar $\hat{S}$ (t) para los tiempos ordenados) puede considerarse lo siguiente:

1a. Supongamos que tenemos $\hat{S}_p(t)$ estimado bajo cierta suposición del Modelo Paramétrico (Valor Extremo, Exponencial, Weibull). ¿ Cual es el criterio para asegurar un buen ajuste a los datos para el modelo considerado?

2a. Suponiendo que se tienen dos o mas modelos que producen un buen ajuste en el punto 1o., ¿ Cuál de los dos (o mas) se ajusta mejor a los datos? (ver Atkinson A.C.,1980) .[17]

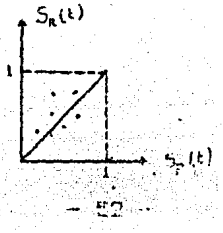
3a. ¿ Existe un criterio no gráfico para asegurar bondad de ajuste en el modelo de Cox ? (ver Schoenfeld D.,1980) [18].

En este capítulo analizaremos la primera pregunta. La segunda y tercera se encuentran resueltas en las referencias antes mencionadas cuyos análisis quedan fuera del objetivo de esta tesis.

4.3.1 Bondad de Ajuste de Modelos

Para asegurar la bondad de ajuste podemos considerar dos formas para lograrlo: La primera consiste en un análisis gráfico, tanto para el caso paramétrico como para el semiparamétrico, para juzgar el ajuste realizado a los datos observados (estimados con Kaplan-Meier) y los ajustados mediante GLIM en el caso paramétrico (Weibull, Exponencial y Valor Extremo) así como en el semi-paramétrico, se tiene lo siguiente:

Si graficamos los valores de la $\hat{S}_K(t)$ contra $\hat{S}_p(t)$ observado (formando así el conjunto de parejas $(\hat{S}_p(t), \hat{S}_K(t))$ que se encuentran entre los intervalos $0 < \hat{S}_K(t) < 1$ y $0 < \hat{S}_p(t) < 1$, resulta así un cuadrado cuyos lados tienen longitud uno y vértices (0,0), (1,1).como en la siguiente figura

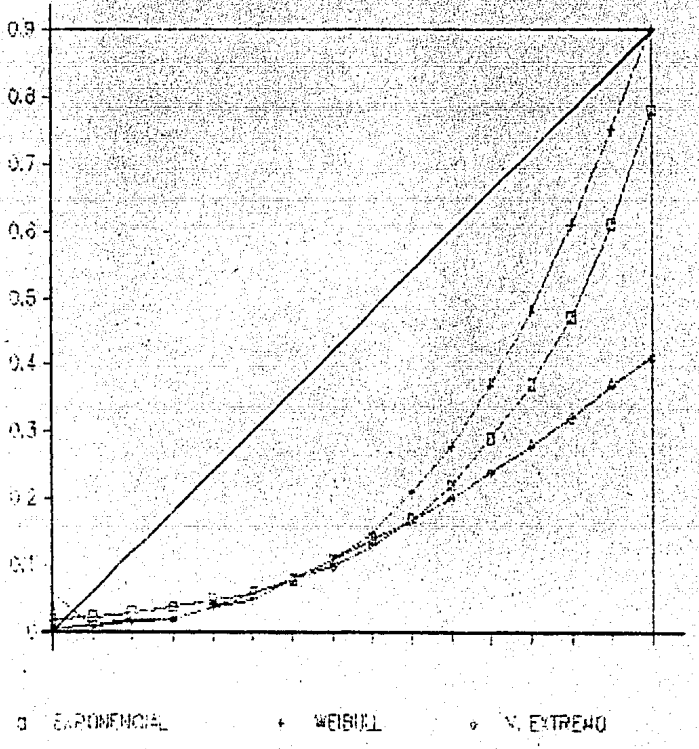


Un ajuste perfecto será cuando la gráfica obtenida pasa sobre la recta que une los puntos (0,0) y (1,1)_.

Un buen ajuste será aquél que permanezca "sufucientemente cerca" de la recta mencionada. Así si graficamos (en una misma gráfica) $\hat{S}_K(t)$, $\hat{S}_P(t)$ para el Valor Extremo, Exponencial y Weibull se tiene a la figura que se muestra en la gráfica 4.7.

grafica 4.7

Datos ajustados de la funcion de supervivencia bajo el modelo Exponencial ,Weibull y Valor Extremo , contra el estimador de Kaplan-Meier.

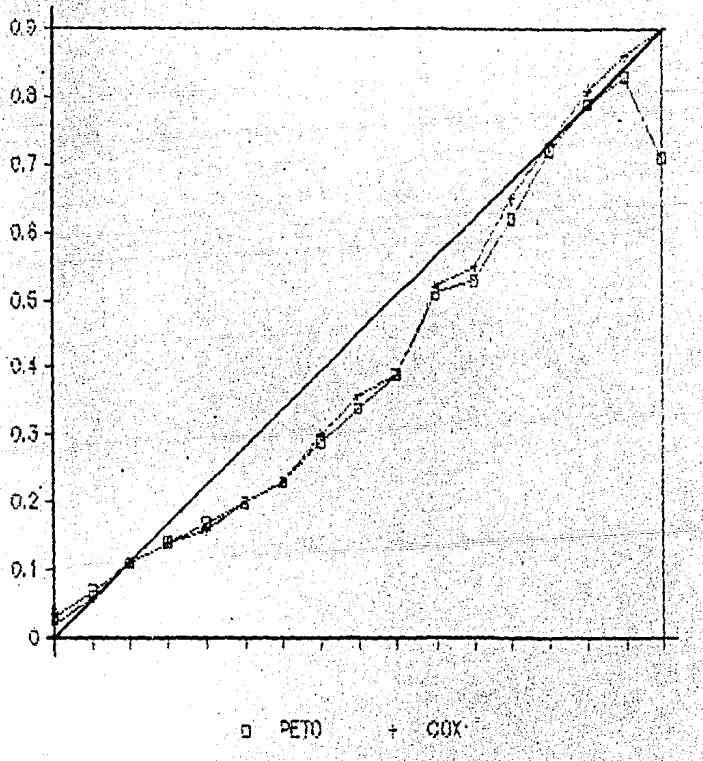


Como puede verse en la figura la gráfica que mejor se ajusta o acerca a la recta es la correspondiente a la distribución Weibull, en seguida la ajustada para la función exponencial y por último para la función Valor Extremo. Por tanto el "mejor" ajuste, en el caso paramétrico, se obtiene para la función Weibull; aunque es notoria la discrepancia que existe con el ajuste "perfecto".

Del mismo modo al hacer esta comparación para los ajustes de Cox y Peto se tiene la gráfica 4.8.

Gráfica 4.8

Datos ajustados de la función de supervivencia bajo el modelo de Cox $\hat{S}_0(t)$ y Peto $\hat{S}_P(t)$ contra el estimador Kaplan-Meier $\hat{S}_K(t)$.



Aquí se observa que el ajuste es bastante bueno relativo al ajuste perfecto.

La segunda consiste en dar una medida cuantitativa del ajuste realizado para cada modelo considerado.

Cuando la información es completa(sin censuras) se usa la prueba de bondad de ajuste X^2.

$$S^2=\sum_{i=1}^{n}\frac{(f_i-e_i)^2}{e_i} \ ,$$

donde $e_i = N(S(t_i) - S(t_{i+1})$,

f_i = frecuencia de sobrevivientes para t,

$N = \sum_{i=1}^{n} f_i$; con K-1-S grados de libertad.

Con s el número de parámetros estimados. Este análisis puede llevarse a cabo solamente para el grupo 1 de los datos que se encuentran en el cuadro 4.1.

Al hacerlo para el modelo paramétrico resulta $S^2 > 100$ con 18 grados de libertad para Weibull en la cual resultan discrepancias significantes bajo la hipótesis nula.

Otro análisis aplicable tanto al caso censurado como no censurado es la estadística no paramétrica de Kolmogorov-Smirnov la cual se usa para el caso de datos no censurados, para probar la hipótesis H_0: $F_{obs.}(t)=F_{ajust.}(t)$ contra la alternativa H_A: $F_{ajust}(t) \neq F_{obs.}(t)$ para N datos observados. La estadística K-S se define como:

$$D_N = \max \left| F_{obs.}(t) - F_{ajust.}(t) \right| .$$

Para esta se tiene que:

si $\sqrt{N}D_N > Y_{1-\alpha}$ se rechaza H_0 ,

si $\sqrt{N}D_N < Y_{1-\alpha}$ se acepta H_0 ,

con $Y_{1-\alpha}$ dados en el siguiente cuadro.[ver Elandt-Jonhson].

Cuadro 4.10
Para diversos α (niveles de significancia).

α	.10	.05	.025	.01
$Y_{1-\alpha}$	1.2238	1.3581	1.4802	1.6276

Por ejemplo para los datos ajustados mediante el modelo Weibull, para el grupo 1 (N=21), D =.36

$$\sqrt{N}\ (D_N)=1.6497 .$$

Como puede verse en el cuadro, para cualquier considerada

$$\sqrt{N}\ (D_N) \quad > Y_{1-\alpha}\ ,$$

por lo tanto se rechaza H_0.

Para el caso del modelo de Cox se tiene para la muestra (1) $\sqrt{N}\ D_N = 0.36$.

Para el cual se acepta H_0 para cualquier nivel de significancia (ver cuadro 4.10).

Para el caso de datos censurados se tiene la siguiente estadística

$$D\ (t) = \max_{t \leq t_c} \left| F_{obs}(t) - F_{ajust.}(t) \right| ,$$

donde t_c es el tiempo de truncamiento del estudio

$$D\ (t\ ) = \max_{t \leq t\phi} \left| F_{obs.}\ (t) - F_{ajust.}(t) \right|$$

$\frac{f}{N} = \phi$, f es el número de muertos en todo el intervalo considerado de tiempo y N son los elementos de la muestra. Para el caso de Weibull se tiene: N=21, D =.34 , f=9

$\sqrt{N}D_N = 1.558$, usando el cuadro 4.11 las discrepancias significativas para =10% son aquellas mayores que =.4286 y por lo tanto se rechaza H_0 , es decir, se tiene un mal ajuste.

En el caso de Cox se tiene: N=21, f=9, D =.08 $\sqrt{N}D_N = .366$, en este caso a cualquier nivel de significancia se acepta H_0 , es decir, se tiene un "buen ajuste" .

Cuadro 4.11

Valores Críticos de $Y_{1-\alpha}$ (ϕ) para distribuciones truncadas del estadístico $\sqrt{N}\ D_N(\phi)$ (ver Elandt-Jonhson].)

	ϕ									
α	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0.10	0.1953	0.2753	0.3360	0.3867	0.4308	0.4703	0.5062	0.5392	0.5699	0.5985
0.05	0.2233	0.3147	0.3839	0.4417	0.4920	0.5569	0.5577	0.6152	0.6500	0.6825
0.025	0.2488	0.3505	0.4276	0.4918	0.5477	0.5975	0.6428	0.6844	0.7230	0.7589
0.01	0.2796	0.3938	0.4803	0.5523	0.6149	0.6707	0.7214	0.7679	0.8110	0.8512
0.005	0.3011	0.4240	0.5171	0.5946	0.6619	0.7219	0.7764	0.8264	0.8726	0.9157
0.001	0.3466	0.4880	0.5950	0.6840	0.7613	0.8303	0.8927	0.9500	1.0029	1.0523

	ϕ									
α	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
0.10	0.5985	0.8155	0.9597	1.0616	1.1334	1.1813	1.2094	1.2216	1.2238	1.2238
0.05	0.6825	0.9268	1.0868	1.1975	1.2731	1.3211	1.3471	1.3568	1.3581	1.3581
0.025	0.7989	1.0282	1.2024	1.3209	1.3997	1.4476	1.4717	1.4794	1.4802	1.4802
0.01	0.8512	1.1505	1.3419	1.4696	1.5520	1.5996	1.6214	1.6272	1.6276	1.6276
0.005	0.9157	1.2361	1.4394	1.5735	1.6582	1.7056	1.7258	1.7306	1.7308	1.7308
0.001	1.0523	1.4171	1.6456	1.7931	1.8828	1.9292	1.9464	1.9494	1.9495	1.9495

Note: [illegible] Y (1 -) = Y

4.3.2 Bondad de ajuste con covariables

Dentro del caso paramétrico y semi-paramétrico,existe también el problema de discernir de entre los modelos, cual de ellos se ajusta mejor a los datos . Los resultados para los datos de Gehan se resumen en el siguiente cuadro.

Cuadro 4.12

devianzas y grados de libertad de los modelos

funcion	A	A+Z	A+Z+T
Exponencial	54.50(41)	38.02(40)	-
Weibull	53.78(41)	34.13(40)	-
Valor Extremo	206.3(41)	180.0(40)	-
Peto	42.85(17)	27.63(16)	27.62(15)
Cox	46.54(29)	30.29(28)	30.27(27)

El modelo A en todos los casos mide los efectos de la variable tiempo como única influencia sobre el fenómeno. El modelo A+Z mide la influencia de la covariable Z (en nuestro caso el grupo) sobre el fenómeno considerado y, por último el modelo A+Z+T mide las posibles discrepancias con el modelo de riesgos proporcionales propuesto por Cox, esto es que las covariables puedan depender del tiempo.

Cuando los modelos son anidados [19] ,existe una prueba de comparación formal que hace uso de las devianzas calculadas para cada modelo;esta es la prueba F de ajuste relativo.

Supongamos que se tienen dos modelos anidados m y m' en el cual m' está anidado en m, sean D y Dm' las devianzas del modelo m y m' consideradas, así como glm y glm' sus grados de libertad respectivos.

Así, si m es el modelo correcto entonces el modelo m' también es correcto. Esta hipótesis es probada comparando la estadística:

$$F=\frac{(Dm' - Dm)/(glm' - glm)}{Dm/glm}$$

con una distribución F con v=glm'-glm grados de libertad en el numerador y m grados de libertad en el denominador.

Valores significativamente grandes de F sugieren que m se ajuste mejor que m'.

Por ejemplo para Weibull se tiene:

$$Dm'=53.78 \text{ y } glm'=41 ,$$

$$Dm=34.3 \text{ y } glm=40.$$

$$F= \frac{(53.78-34.13)/(41-40)}{34.13/\ 40} = \frac{(19.65)(40)}{34.13} ,$$

$$F=23.03$$

con $v=1$

$g=40$,

el valor en tablas al 5% es F =4.08, dado que el valor de F es significativamente mayor que el de tablas, se tiene que A+Z se ajusta mejor que A en este caso.

Para el caso de Cox, si tomamos:

$$m=A+Z+T$$

$$m'=A+Z ,$$

se tiene:

$$Dm=30.27 \quad glm=27$$
$$Dm'=30.29 \quad glm'=28 ,$$

$$F=\frac{(30.29-30.27)/(28-27)}{30.27/27} = \frac{(0.02)27}{30.27} = .017 ,$$

$$v=1$$

$$glm=27$$

$$F_T=4.21 \text{ al } 5\% ,$$

como $F << F_T$,

no parece que el modelo A+Z+T sea mejor que el modelo A+Z. Sin embargo, sin definimos

$$m = A + Z$$

$$m' = A \quad ,$$

$$Dm = 30.29 \qquad glm = 28$$

$$Dm' = 46.54 \qquad glm' = 29 \quad ,$$

$$F = \frac{(46.54 - 30.29)/(29-28)}{(30.29)/28} = 16.25 \quad ,$$

$$F = 4.19 \text{ al } 5\%$$

$$v = 1 \quad , \quad glm = 28 \quad ,$$

el ajuste es mejor para el modelo A+Z que para el modelo A.

CONCLUSIONES

Se han planteado en este trabajo dos formas generales de analizar los fenomenos de supervivencia que son: el modelo paramétrico y el modelo semi-paramétrico. A su vez dentro del modelo paramétrico se consideraron dos formas de abordar el problema de cálculo de los parámetros, el modelo paramétrico y el modelo paramétrico general. El primero aborda el análisis de algunos casos particulares haciendo la estimación de los parámetros de forma y de las covariables separadamente.

De este modo el modelo paramétrico general presenta ciertas ventajas dado que reune en un sólo modelo varias funciones paramétricas asi como la estimacion de los parametros desconocicos en cada función utilizada; calculando al mismo tiempo los parámetros desconocidos relacionados con las covariables.

Las limitaciones que presentan los modelos paramétricos son, principalmente, que el número de funciones posibles a usar es pequeño considerando las posibilidades de distribución de los datos de supervivencia y que su uso depende de la posibilidad de ajuste de los datos a una de ellas.

Como hemos visto su implementación en GLIM es posible. Sin embargo, al analizar los datos de Gehan,se observa que el ajuste de los modelos considerados es bastante pobre para esos datos. Esto significa que el poder de predicción y explicación del modelo será también malo en general.

El modelo semi-paramétrico, fundamentalmente el modelo de Cox, tiene amplias ventajas sobre los modelos paramétricos; entre ellas podemos destacar las siguentes:

a) No es necesario conocer explícitamente la función de riesgo fundamental $\lambda_0(t)$ y sólo se supone de ella que es mayor que cero y bien comportada.

b) La información dada por el orden estadístico de los tiempos arroja información acerca del comportamento de la función de supervivencia y por otro lado permite la estimación de los parámetros vía la función de verosimilitud parcial.

c) Su implementación es posible en GLIM (vía modelo Poisson).

d) Existe una aproximación propuesta por Peto (para tratar casos de observaciones coincidentes o empates) que es mas conveniente desde el punto de vista de cálculo.

e) El ajuste del modelo a datos de supervivencia es bastante bueno considerando los resultados que se obtienen al aplicar el modelo a casos particulares. Por tanto el poder de predicción es bastante bueno así como el de explicación del modelo con base en las covariables consideradas.

Una desventaja importante es que el cálculo numérico es bastante engorroso para el modelo de Cox, aún en la aproximación de Peto. Recientemente se ha elaborado un algoritmo (del tipo EM; ver Clayton y Cuzick,1985)) que sirve para calcular con más facilidad los parámetros del modelo, aplicable cuando las covariables son fijas. Desgraciadamente no es posible usarlo cuando las covariables dependen del tiempo.

Con respecto a la bondad de ajuste en el caso paramétrico hemos visto que existen estadísticas como la χ^2 (en el caso de datos no censurados) y la de Kolmogorov-Smirnov (en datos censurados) que sirven para medir la bondad de ajuste de un modelo con respecto a otro.

La bondad de ajuste del modelo de Cox no puede ser evaluado por referencia a tablas . Esto es debido a que el número de parámetros en el modelo saturado se incrementa con el número de observaciones, violando asi las suposiciones fundamentales de la justificación asintótica de la prueba .

Lo mismo sucede para el tratamiento de Peto, a menos que se suponga que ni el número de grupos ni el número de tiempos de muerte distintos crece sin límite con el número de observaciones.
En tal caso las comparaciones de los modelos puede llevarse a cabo mediante tablas .

Existen algunos enfoques más elaborados para asegurar la bondad de ajuste del modelo de riesgos de Cox como el de Schoenfeld. Sin embargo su análisis rebasa los objetivos de esta tesis.

APENDICE A

PROGRAMAS DE AITKIN Y CLAYTON

Los métodos descritos por Aitkin y Clayton para ajuste de los modelos de regresión paramétrica a datos de supervivencia consista en un algoritmo que se lleva a cabo generalmente mediante dos pasos.

Primeramente mostraremos el archivo de datos de supervivencia para el ajuste de la función exponencial y Weibull, con el ejemplo que se encuentra en el capítulo IV. La primera columna son los tiempos de supervivencia; la segunda columna de ceros y unos es la asignación de censura o falla a cada individuo; para la muestra uno y posteriormente para la muestra cero,

Muestra uno		Muestra cero	
1	1	6	1
1	1	6	1
2	1	6	1
2	1	6	0
3	1	7	1
4	1	9	0
4	1	10	1
5	1	10	0
5	1	11	0
8	1	13	1
8	1	16	1
8	1	17	0
8	1	19	0
11	1	20	0
11	1	22	1
12	1	23	1
12	1	25	0
15	1	32	0
17	1	32	0
22	1	34	0
23	1	35	0
		1	0

Ajuste de la función Exponencial.

El ajuste de este modelo se lleva a cabo en un solo paso, pues la media sólo involucra parámetros de escala.

A los tiempos se les denota con la letra T y a las censuras con la letra C, la variables explicativa, que en este caso es la identificación del grupo, la denotaremos con la letra X.

```
R$SERVICIO/GLIM;FILE FILE1(DISK,TITLE=GEHAN,FILETYPE=7);
#RUNNING 8267

$?

GLIM 3.11 (C)1977 ROYAL STATISTICAL SOCIETY, LONDON

$UNITS 42 $DATA T C $DIN 1 50 $

$CA X =1.5-%GL(2,21) $

$CA LT=%LOG(T) $OFF LT $
$YVAR C $ERR P $
$F$

        SCALED
 CYCLE  DEVIANCE       DF
    5    54.50         41
$FIT + X $

        SCALED
 CYCLE  DEVIANCE       DF
    4    38.02         40

$DIS ME $

Y-VARIATE  C
ERROR POISSON  LINK LOG
OFFSET LT

LINEAR PREDICTOR
%GM X

       ESTIMATE       S.E.       PARAMETER
   1   -2.923       0.1979       %GM
   2    1.526       0.3958       X
  SCALE PARAMETER TAKEN AS       1.000
```

Ajuste de la funcion Weibull.

En este modelo intervienen los dos pasos del algoritmo, dado que la media no sólo involucra parámetros β, sino también un parámetro de forma. El primer paso consiste en maximizar α manteniendo constante β. El segundo mantiene fija β maximizando la verosimilitud para α, este proceso itera hasta la convergencia.

Para llevar a cabo el ajuste este utiliza una subrutina propuesta por Aitkin y Clayton (1980). De igual manera que en el ajuste exponencial los tiempos se denotan con la letra T y las censuras con la letra C, y la variable explicativa es X.

```
R$SERVICIO/GLIM;FILE FILE1(DISK,TITLE=GEHAN,FILETYPE=7)
#RUNNING 3168
#3168 WARNING: THIS 31 CODE FILE CAN NOT BE RUN ON THE
#?

GLIM 3.11 (C)1977 ROYAL STATISTICAL SOCIETY, LONDON
$UNITS 42 $DATA T C $DIN 1 50 $

$CAL X=1.5-%GL(2,21)$

$INPUT 2 WEIBULL$

$CA %N=1$

$M MODEL X $E

$ARG WEIBULL T C %N

$USE WEIBULL
```

Subrutina que se emplea para el ajuste de la función Weibull.

```
3200 $M MESS $PR '***WARNING!
3300 STANDARD ERRORS OF ESTIMATES GIVEN BELOW ARE UNDERESTIMATED!
3400 (AS THEY DO NOT ALLOW FOR THE ESTIMATION OF THE SHAPE PARAMETER)!
3500 SEE AITKIN AND CLAYTON,APPL. STATS.2,1980 FOR CORRECT PROCEDURE'!
3600 $$E
3700 !
3800 !
3900 $M MOD1 $PR 'EXPONENTIAL FIT' $$E
4000 !
4100 $M MOD2 $PR 'WEIBULL FIT' $$E
4200 !
4300 !
4400 $M WB1 !
4500 $CA %W=%W-1 ! %W=%GT(%W,0)*%W !
4600 ! ZZ1=%A*%1 $OF ZZ1 !
4700 $OU $F *MODEL $REC $OU 6!
4800 $CA %D=%Z - 2*%CU(%2*%LOG(%A*%FV)-%FV) !
4900 ! %X=1+%NE(%A,1) !
5000 $SW %X MOD1 MOD2 !
5100 $CA %X=%DF-%NE(%A,1) !                          CALCULATE CORRECT DF
5200 $PR ! ' DEVIANCE   SHAPE    DF'!
5300 ! '             PARAMETER'  !
5400 !%5 %D %A *-1 %X !! !
5500 $CA %V=%EQ(%3,0) $EX %V !
5600 $CA %X=%CU(%1*(%FV-%2)); %X=0.5*(%A-%Y/%X)!     NEW INCREMENT FOR SHAPE
5700 ! %V=%LE(-.00001,%X)*%GE(.00001,%X) $EX %V!     TEST FOR CONVERGENCE
5800 $CA %A=%A-%X $$E  !                               UPDATE SHAPE PARAMET
5900 !
6000 !
6100 $M WEIBULL !
6200 $CA ZZ2=%LOG(%1+0.5*%EQ(%1,0))!                 ZZ2 IS TRANSFORMED SURVI
6300 ! %A=1 !                                         INITIAL VALUE OF SHAPE
6400 ! %Y=%CU(%2) !
6500 ! %Z=2*%CU(ZZ2*%2) !
6600 ! %W=15 !                                       MAX NO. OF ITERATIONS
6700 $Y %2 $ER P $CYC !
6800 $ARG WB1 ZZ2 %2 %3 $WH %W WB1 !                  USE WB1 UNTIL CONVERGE
6900 $SW %3 MESS !
7000 $D ERT !
7100 $DEL ZZ1 ZZ2 !
7200 $$E
7300 !
7400 !
7500 $M RESPLOTS !
7600 $SW %3 %2 %1 : %1 !
7700 $CA ZZ1=(%CU(1)-%EXP(-%1)*%CU((1-%2)*%FXP(%1)))/(%NU+1) !
7800 : ZZ2= -%LOG(1-ZZ1).
7900 $P ZZ2 %1 $ !
8000 $CA ZZ1=ZZ1 + %EQ(ZZ1,0)*0.000001 !
8100 $CA ZZ1=%ANO(1-ZZ1) ! %1=%ANO(%EXP(-%1)) !
8200 $P ZZ1 %1 !
8300 $DEL ZZ1 ZZ2 !
8400 $E
8500 !
```

Estimadores de los parámetros de la

funcion Weibull

```
WEIBULL FIT

 DEVIANCE    SHAPE      DF
             PARAMETER
   213.16    1.3658     39.


***WARNING
STANDARD ERRORS OF ESTIMATES GIVEN BELOW ARE UNDERESTIMATED
(AS THEY DO NOT ALLOW FOR THE ESTIMATION OF THE SHAPE PARAMETER)
SEE AITKIN AND CLAYTON,APPL. STATS,2,1980 FOR CORRECT PROCEDURE
        ESTIMATE        S.E.        PARAMETER
    1    -3.936        0.1992        %GM
    2     1.731        0.3984        X
   SCALE PARAMETER TAKEN AS          1.000
```

APENDICE B

PROGRAMAS DE ROGER Y PEACOCK

Los autores mediante unas pequeñas modificaciones en las declaraciones internas del paquete hacen que los parámetros α y β se ajusten simultáneamente. En la función de verosimilitud unifican el ajuste de la función Weibull y Valor Extremo en un solo programa. La única diferencia entre dichas funciones es la escala de la variable U, que en el caso Weibull es U=log t.

```
R$SERVICIO/GLIM;FILE FILE1(DISK,TITLE=[illegible],FILETYPE=7),FILE2(DISK,TI
#QUEUED INPUT DISCARDED
#RUNNING 9323
#?

GLIM 3.11 (C)1977 ROYAL STATISTICAL SOCIETY, LONDON
$INPUT 2 SWEIBULL $
$UNITS 43 $DAT T C $DIN 1 10 $
$CAL X=1.5-%GL(2,21) :X(%NU)=0
$CAL U=%LOG(T)
$USE SETW
$FIT +X $
$DISP E M R $

        SCALED
 CYCLE  DEVIANCE        DF
   4     53.78          41

      ESTIMATE       S.E.        PARAMETER
   1  -3.291        0.[illegible]      GM
   2   1.141        0.1694       U
   SCALE PARAMETER TAKEN AS    1.000

        SCALED
 CYCLE  DEVIANCE        DF
   3     3[illegible]          40

      ESTIMATE       S.E.        PARAMETER
   1  -3.9[illegible]       0.5951       GM
   2   1.[illegible]        0.[illegible]       U
   3   1.[illegible]        0.[illegible]41       X
   SCALE PARAMETER TAKEN AS    1.000

? [illegible]RIATE [illegible]PAGE.
```

Programa que ajusta los datos a la

distribución Valor Extremo

```
#
R$SERVICIO/GLIM;FILE FILE1(DISK,TITLE=DEHAN,FILETYPE=71,FILE2(DISK,TITLE
$RUNNING 8498
#?

GLIM 3.11 (C)1977 ROYAL STATISTICAL SOCIETY, LONDON
$INPUT 2 SWEIBULL $
$UNITS 43 $DAT T C $DIN 1 50 $
$CAL X=1.5-%GL(2,21) /%(%NU)=0
$CAL U=T
$USE SETW
$FIT +X $
$DIS E M R $

        SCALED
 CYCLE  DEVIANCE      DF
    4     206.3       41

       ESTIMATE       S.E.       PARAMETER
    1   -1.757       0.3030       CH
    2   .8326E-01    .1123E-01    U-
   SCALE PARAMETER TAKEN AS        1.000

        SCALED
 CYCLE  DEVIANCE      DF
    4     180.0       40

       ESTIMATE       S.E.       PARAMETER
    1   -2.395       0.3827       CH
    2   0.1154       .1588E-01    U
    3    2.172       0.4034       X
   SCALE PARAMETER TAKEN AS        1.000

Y-VARIATE  C
ERROR OWN      LINK OWN
MW1
MW2
MW3
MW4

LINEAR PREDICTOR
CM U X
```

S W E I B U L L

Subrutina(macros).Estas son las modificaciones hechas en las declaraciones internas del paquete. Empleada en el programa que ajusta a la distribución Weibull y Valor Extremo.

La macro(SETW) ajusta primeramente el modelo nulo o sea sin covariables.

```
L SURVIVAL
$FILE (ISSC)SURVIVAL ON IIMAS
100 $SUBFILE SWEIBULL
200 !
300 !
400 $MAC MW1 !
500 $CAL %LP=%IF(%LT(%LP,78),%LP,78) !
600 : %LP=%IF(%GT(%LP,-78),%LP,-78) !
700 : %FV=%EXP(%LP) : %FV(%NU)=%W/2 !
800 $END !
900 $MAC MW2 !
1000 $CAL %DR=1/%FV :%DR(%NU)=%LP(%NU)/%FV(%NU) !
1100 $END !
1200 $MAC MW3 !
1300 $CAL %VA=%FV : %VA(%NU)=%W/4 !
1400 $END !
1500 $MAC MW4 !
1600 $CAL %DI=2*(%FV-C*(%LP+1)) : %DI(%NU)=-2*%W*%LOG(%LP(%NU)) !
1700 $END !
1800 $MAC SETW !
1900 $CAL GM=1 : GM(%NU)=0 : U(%NU)=1 :C(%NU)=0 : %W=%CU(C) !
2000 : C(%NU)=%W !
2100 $YVAR C !
2200 $OWN MW1 MW2 MW3 MW4 !
2300 $SCALE 1$ !
2400 $CAL %LP=%LOG(C*0.8+0.1) : %LP(%NU)=0.5 !
2500 $FIT GM - %GM + U !
2600 $DIS E $ !
2700 $END !
```

APENDICE C

AJUSTE DE MODELOS SEMI-PARAMETRICOS

C.1 APROXIMACION DE PETO

Continuaremos con el ejemplo de los datos de Gehan(1965), en los cuales se tienen pacientes con leucemia, divididos para su estudio en dos grupos; uno de ellos el grupo cero ha sido tratado con una droga (6-mp) y la muestra uno con placebo, algunas observaciones son censuradas (*) (ver cuadro 4.1 cap. 4).

Las observaciones no censuradas cubren 17 intervalos , los grupos quedan divididos en N_{h0} y N_{h1} respectivamente, donde N_{h0} es el numero de tiempos de supervivencia en la muestra "0" los cuales son mayores o iguales a t y m_h es la multiplicidad de muertes al tiempo t en el grupo, lo mismo sucede para el otro grupo,los datos se muestran en el siguiente cuadro.

t	N	m	N	m
1	21	0	21	2
2	21	0	19	2
3	21	0	17	1
4	21	0	16	2
5	21	0	14	2
6	21	3	12	0
7	17	1	12	0
8	16	0	12	4
9	16	1	8	0
10	15	1	8	0
11	13	0	8	2
12	12	0	6	2
13	12	1	4	0
15	11	0	4	1
16	11	1	3	0
17	10	0	3	1
22	7	1	2	1
23	6	1	1	1

El ajuste del modelo de Peto vía modelo de Poisson para fallas múltiples es como sigue : Para cada t_h se tienen X_{h0} y X_{h1} como observaciones de Poisson con parámetros

$$\mu_{h0} = N_{h0} \exp(\hat{\alpha}_h + Z_{h0}\hat{\beta}),$$

y

$$\mu_{h1} = N_{h1} \exp(\hat{\alpha}_h + Z_{h1}\hat{\beta}),$$

donde

$Z_{h0}=(0,0)$ y $Z_{h1}=(1,t-10)$, en el que se introduce el tiempo como una posibilidad de no proporcionalidad para el muestra uno y t representa el inicio del tratamiento. Los parámetros α_h se ajustan como un factor con 17 niveles, uno para cada t_h (h=1,2,...17), como un factor con con dos niveles ($\beta_1=0$ para la muestra "0" y $\beta_1=1$ para la muestra "1", se ajusta con regresión simple y se usa $\ln N_{hj}$ (j=0,1) como un OFFSET.

Programa que ajusta los datos de supervivencia mediante la aproximación de Peto.

```
R$SERVICIO/GLIM;FILE FILE1(DISK,TITLE=WHI/HEAD,FILETYPE=7);
$RUNNING 8101

$?

GLIM 3.11 (C)1977 ROYAL STATISTICAL SOCIETY, LONDON
$UNITS 34 $DATA N X T $DIN 1 50 $FAC Z 2 $CAL Z=%GL(2,17) $

$FAC A 17 $CAL A=%GL(17,1) $CAL N1=%LOG(N) $OFF N1 $YVAR X $ERR P $

$VAR 17 IND $CA IND=%GL(17,1) $CA T=T-10 $CA T(IND)=0 $

$FIT A -%GM $

      SCALED
CYCLE DEVIANCE      DF
   4    42.85       17
$FIT + Z $

      SCALED
CYCLE DEVIANCE      DF
   5    27.63       16
$FIT + T $

      SCALED
CYCLE DEVIANCE      DF
   5    27.62       15
```

Parámetros α_h, que se emplean en la construcción de la función de Supervivencia (Aproximación de Peto).

```
$DIS ME $

Y-VARIATE  X
ERROR POISSON  LINK LOG
OFFSET N1

LINEAR PREDICTOR
A Z
```

	ESTIMATE	S.E.	PARAMETER
1	-8.976	1.231	A(1)
2	-8.807	1.223	A(2)
3	-4.678	1.058	A(3)
4	-8.524	1.209	A(4)
5	-8.311	1.197	A(5)
6	-11.28	1.366	A(6)
7	-4.358	1.055	A(7)
8	-13.86	1.651	A(8)
9	-4.021	1.048	A(9)
10	-7.196	1.187	A(10)
11	-6.712	1.166	A(11)
12	-3.477	1.036	A(12)
13	-3.446	1.038	A(13)
14	-3.269	1.031	A(14)
15	-3.230	1.033	A(15)
16	-4.773	1.096	A(16)
17	-3.819	1.041	A(17)
18	1.628	0.4320	Z
SCALE PARAMETER TAKEN AS			1.000

C.2 APROXIMACION DE COX

De la misma manera que la aproximación de Peto en este caso los vectores correspondientes a los pacientes de la muestra cero son $Z_{h0}=(0,0)$ y para la muestra uno $Z_{h1}=(1,T-10)$. De esta forma quedan asignadas las funciones de riesgo $\lambda_0(t)$ y $\lambda_0(t)\exp\{\beta_1+\beta_2(t-10)\}$ respectivamente. El parámetro β_1 mide así las diferencias entre las tasas de falla entre los dos grupos y β_2 mide esa diferencia en el tiempo.

Para el tratamiento de multiplicidad en esta aproximación consideremos el caso $t_5=5$, existen en este tiempo $N_5=35$ pacientes en riesgo; $N_{5,0}=21$ en la muestra cero y $N_{5,1}=14$ en la muestra 1. También existen $M_4=\binom{35}{2}=595$ posibles pares de pacientes en riesgo, formando 3 grupos diferentes. En primer grupo consiste de $M_3=\binom{21}{2}=210$ pares de pacientes tomados ambos de la muestra cero. El vector de covariables para esos pares es $S_{5,0}=(0,0)$ que es la suma de las covariables de dos individuos en la muestra cero.

Ademas existen $M_{5,2}=21\times14=294$ pares de pacientes tomados uno de cada muestra y tienen vector de covariables $S=(1,\ t-10)$. Finalmente $M_{5,2}=\binom{15}{2}=105$ pares de pacientes tomados de la muestra 1 con vector $S_{5,1}=(2,\ 2t-20)$

Las variables Poisson que corresponden a este tiempo de muerte son $X_{5,0}$, $X_{5,1}$ y $X_{5,2}$ a ellos les corresponden los parámetros:

$$\mu_{5,0}=M_{5,0}\exp(\alpha_5),$$

$$\mu_{5,1}=M_{5,1}\exp(\alpha_5+\beta_1-5\beta_2),$$

$$\mu_{5,2}=M_{5,2}\exp(\alpha_5+2\beta_1+10\beta_2).$$

Cuyos valores son 0,0 y 1 respectivamente dado que ambas muertes caen en el grupo 1. Así se lleva a cabo para cada t_h.

En total se obtienen 46 observaciones poisson aquí de nuevo α_h se ajusta como un parametro con 17 niveles, β_1 como un factor con 2 y β_2 como un cosficiente a calcular, log M_{hj} son usadas como un offset.

La forma en que se arreglan los datos para ejecutar la corrida en GLIM es como sigue:

	N	X	Z (1,t-10)	A
21				
2	210	0	(0,0)	1
21 21				
1 1	441	0	(1,-9)	1
21				
2	210	1	(2,-18)	1
21				
2	210	0	(0,0)	2
21 19				
1 1	399	0	(1,-8)	2
19				
2	171	1	(2,-16)	2
21				
1	21	0	(0,0)	3
17				
1	17	1	(1,-7)	3
21				
1	210	0	(0,0)	4
21 16				
1 1	336	0	(1,-6)	4
16				
2	120	1	(2,-12)	4
21				
2	210	0	(0,0)	5
21 14				
1 1	294	0	(1,-5)	5
14				
2	91	1	(2,-10)	5
21				
3	1330	1	(0,0)	6
21 12				
2 1	2520	0	(1,-4)	6
21 12				
1 2	1386	0	(2,-8)	6
12				
3	220	0	(3,-12)	6
17				
1	17	1	(0,0)	7
12				
1	12	0	(1,-3)	7
16				
4	1820	0	(0,0)	8
16 12				
3 1	6720	0	(1,-2)	8
16 12				
2 2	7920	0	(2,-4)	8
16 12				
1 3	3520	0	(3,-6)	8

12				
4				
15	495	1	(4,-8)	8
1				
8	15	1	(0,0)	9
1				
13	8	0	(1,0)	9
2				
13 8	78	0	(0,0)	10
1 1				
8	104	0	(1,1)	10
2				
12	28	1	(2,2)	10
2				
12 6	66	0	(0,0)	11
1 1				
6	72	0	(1,2)	11
2				
12	15	1	(2,4)	11
1				
4	12	1	(0,0)	12
1				
11	4	0	(1,3)	12
1				
4	11	0	(0,0)	13
1				
11	4	1	(1,5)	13
1				
3	11	1	(0,0)	14
1				
10	3	0	(1,6)	14
1				
3	10	0	(0,0)	15
1				
7	3	1	(1,7)	15
2				
7 2	21	0	(0,0)	16
1 1				
2	14	1	(1,12)	16
2				
6	1	0	(2,24)	16
2				
6 1	15	0	(0,0)	17
1 1				
	6	1	(1,13)	17

Programa que ajusta los datos de supervivencia mediante la aproximación de Cox.

```
R$SERVICIO/GLIM;FILE FILE1(DISK,TITLE=COX1,FILETYPE=7);
#RUNNING 7180

$?

GLIM 3.11 (C)1977 ROYAL STATISTICAL SOCIETY, LONDON


$UNITS 46 $DATA N X Z T A $DIN 1 50 $FAC A 17 $

$CAL N1=%LOG(N) $OFF N1 $

$YVAR X $ERR P $

$FIT A - %GM $

         SCALED
 CYCLE   DEVIANCE        DF
    5      46.54         29

$FIT +Z $

         SCALED
 CYCLE   DEVIANCE        DF
    5      30.29         28
$FIT + T $

         SCALED
 CYCLE   DEVIANCE        DF
    5      30.27         27
```

Parámetros que se emplean para la construcción de la función de Supervivencia (aproximacion de Cox).

```
$DIS M E $

Y-VARIATE  X
ERROR POISSON  LINK LOG
OFFSET N1

LINEAR PREDICTOR
A Z
```

	ESTIMATE	S.E.	PARAMETER
1	-4.060	0.7826	A(1)
2	-3.979	0.7800	A(2)
3	-4.584	1.050	A(3)
4	-3.843	0.7751	A(4)
5	-3.742	0.7711	A(5)
6	-3.223	0.6484	A(6)
7	-4.267	1.047	A(7)
8	-2.866	0.5917	A(8)
9	-3.935	1.041	A(9)
10	-3.202	0.7686	A(10)
11	-2.974	0.7620	A(11)
12	-3.404	1.030	A(12)
13	-3.370	1.032	A(13)
14	-3.201	1.025	A(14)
15	-3.160	1.027	A(15)
16	-2.082	0.7438	A(16)
17	-1.660	0.7287	A(17)
18	1.509	0.4096	Z(2)
SCALE PARAMETER TAKEN AS		1.000	

NOTAS

1.- De aquí en adelante usaremos el término "individuo" para los elementos de la muestra y "falla" al evento que se presenta al tiempo t

2.- Para la construcción empírica de la función de supervivencia, ver el capítulo 5 Elandt-Johnson, R.C. et. al. (1980).

3.- En la literatura referente al fenómeno de supervivencia el término de "tasa instantánea de falla" (Gross, et.al.,1975) "función de riesgo" (Epstein et.al. 1953) "intensidad de riesgo"(Chiang 1968) y " Fuerza de mortalidad" (Jordan 1952) son usados indistintamente para (t).

4.- El uso de la función $g(Z\beta)=\exp(Z\beta)$ tiene amplia aplicación en el estudio del fenómeno de supervivencia, sin embargo, no existe ninguna razón del tipo teórico que obligue a su uso. Su uso, en el análisis de supervivencia, es equivalente al uso de la función Gaussiana en estadística.

5.- La función liga relaciona al predictor lineal "η" al valor esperado μ del dato y. En los modelos lineales clásicos estos son idénticos y la liga identidad es "sensible",en el sentido que, ambos μ y η pueden tomar cualquier valor real. Sin embargo, cuando se tienen datos que involucran la función Poisson es necesario que μ sea positiva y la liga no sería apropiada dado que puede ser negativa. Cuando los modelos se basan en la independencia de probabilidades, estos llevan a considerar efectos multiplicativos y así a una funcion liga $\eta=\ln\mu$ o $\mu=e^{\eta}$ Asi efectos aditivos que contribuyen a η se vuelven efectos multiplicativos en μ. Por otro lado, la distribucion binomial tenemos que $0<\mu<1$ y la liga debe mapear el intervalo (0,1) en todo el eje real.

6.- Una matriz es de rango completo cuando el sistema asociado con ella tiene una solución única, esto sucede si su sistema homogéneo asociado tiene como solución única el vector nulo.

6'. En algunos casos, es necesario ajustar un modelo donde las β_j son fijas en el predictor lineal en el proceso de cálculo. En general, si un subconjunto de las β_j son fijas, la suma de sus contribuciones al predictor lineal η es llamada un offset ; tal que:

$$\eta_i = \text{offset} + \sum_j z_{ij}\beta .$$

7.- Para el caso exponencial y Weibull con datos censurados y sin variables concomitantes pueden hallarse expresiones cerradas para los valores de los parámetros.

Por ejemplo si

$$S(t) = EXP(-\rho t) ,$$

$$\hat{\beta} = \frac{d}{\sum t_i} \quad ; \quad t = \text{suma sobre los censurados y no censurados}$$

Para el caso de Weibul

$$\lambda(t_i) = k\rho(\rho t)^{k-1} \quad \text{se obtiene:}$$

$$\rho = \left(\frac{d}{\sum t_i^k}\right)^{1/k} \quad \text{y} \quad \frac{\sum x_i^k \ln x_i}{\sum x_i^k} = \frac{1 + \sum_u \left(\frac{\ln t_i}{d}\right)}{k} ,$$

u = no censurados

y pueden hallarse k y ρ de estas ecuaciones.

8.- Dado $\mu_i = t_i^{\alpha} \exp z\beta$ para el caso de Weibull y

$\Xi = n \ln\alpha + \sum (\delta_i \ln\mu_i - \mu_i)$ se tiene $\mu_i = \Delta(t_i)\exp(z\beta)$, donde $\Delta(t) = t^{\alpha}$

$$\frac{\partial \Xi}{\partial \alpha} = \frac{r}{\alpha} + \sum \left\{ \delta_i \left(\frac{1}{\mu_i} \frac{\partial \mu_i}{\partial \alpha} \right) - \frac{\partial \mu}{\partial \alpha} \right\} ,$$

además

$$\frac{\partial \mu_i}{\partial \alpha} = \frac{\partial}{\partial \alpha} \left\{ \exp(z\beta) \cdot \exp[\alpha \ln t_i] \right\}$$

$$= \exp(z\beta) \ln t_i \cdot \exp\{\alpha \ln t_i\} .$$

$$\frac{\partial \mu_i}{\partial \alpha} = (\ln t_i)\, t_i^{\alpha} \exp(z\beta) = \mu_i \ln t_i ,$$

$$\frac{\partial \Xi}{\partial \alpha} = \frac{r}{\alpha} + \sum \delta_i \ln t_i - \mu_i \ln t_i$$

$$\frac{\partial \Xi}{\partial \alpha} = \frac{r}{\alpha_i} + \sum (\delta_i - \mu_i) \ln t_i .$$

9.- Bennet y Whitehead en su artículo muestran un programa que ajusta la distribución logística.

10.- En este caso el término de verosimilitud es debido a que, los términos que determinaron cuales individuos deben ser censurados de entre los supervivientes han sido omitidos de cada conjunto de riesgo. Dado que el mecanismo de censura no depende de β, esos términos no dependerán funcionalmente de β y pueden ser ignorados para propósitos de inferencia sobre β .

11.- En lo sucesivo, en esta parte abreviaremos $\lambda(k)=\exp(z_k\beta)$

$$K_{i(k)} = \sum_{k \in R(\tau_j)} \exp(z_k \beta)$$

12.- Debido a la dependencia sobre la multiplicidad g , el producto de todos esos términos del tipo no es una verosimilitud marginal completa de los rangos es necesario justificar la teoria asintotica la verosimilitud parcial.

13.- El tratamiento del modelo de riesgos de Cox lo realiza Kalbfleisch y Prentice en su artículo para mayores detalles ver referencia.

14.-En este caso $S_0(t)$ corresponde con el valor de Z_i en la covariable igual a cero ,con $Z_0=1$ y $S(t)$ con el valor de Z_0 en la covariable igual a uno con $Z_1=1$.

15.- Para una distribucon discreta se define
f_j = atomo de probabilidad al tiempo t_j

$$\lambda(t)=\sum_j \lambda_j \delta(t-a) \quad \text{donde} \quad \lambda_j = \frac{f_j}{S(t_j)}$$

y $\delta(x)$ es la delta de Dirac .

$$S(t)=\prod_{t_j<t}(1-\lambda_j) \qquad y \qquad \Lambda(t)=\sum_{t_i<t}\log(1-\lambda_j)$$

y así $S(t)=\exp(-\Lambda(t))$

si λ_j es pequeno $(\lambda_j << 1)$,

$$\Lambda(t) \cong \sum_{t_j<t} \lambda_j$$

$$S(t)=\exp\left(-\sum_{t_j<t}\lambda_j\right) .$$

16.- Esta aproximación es propuesta por Whitehead (1980) para la expresión de S(t) en el modelo de Cox.

17.- Atkinson en su artículo "Un método para discriminación entre modelos" plantea: Si se tiene
1o.) un vector de variables $\underset{\sim}{Y} = (y_1, \ldots y_n)$ conocidas
2o.) un conjunto de f.d.p. $f_1(\underset{\sim}{Y}, \theta_1)$ y $f_2(\underset{\sim}{Y}, \theta_2)$ bajo consideracion, que pretendan describir los datos, ¿Cuál de los modelos es el mas apropiado?

El problema consiste en elaborar una estadística que discrimine entre los modelos propuestos.

La idea consiste en combinar las dos hipótesis (nula y alternativa) en un modelo general. Considerando las f.d.p. por ejemplo proporcional a:

$$\left\{f_1(\underset{\sim}{Y}, \theta_1)\right\}^{\lambda} \left\{f_2(\underset{\sim}{Y}, \theta_2)\right\}^{1-\lambda},$$

asi, si λ se aproxima a 1 se podra decir que el modelo 1 es mas apropiado que el modelo dos y si $\lambda = 1/2$ serán igualmente válidos.

18.- Schoenfeld plantea un análisis de bondad de ajuste, para el modelo de regresión de "riesgos proporcionales", del tipo Ji-cuadrada en el cual involucra datos censurados y covariables. Su análisis queda mas allá del objetivo de esta tesis.

19.- Se dice que el modelo m y m' son anidados si m' se puede obtener de m haciendo que un subconjunto de variables distintas de cero , tomen el valor de cero.

REFERENCIAS

1.- Aitkin, M. y Clayton, D.(1980).The Fitting of Exponential,Weibull and Extreme Value Distributions to Complex Censored Survival Data using GLIM. Appl. Statist., 29,156-163

2.- Atkinson, A.C.(1980) A note on the generalized information criterion for choise of a model,Biometrika , 7,2,pag.413-18

3.- Aranda-Ordaz,F.J.(1983) An Extension of the Proportional-Hazards Model for Grouped Data. Biometrics 39,109-117.

4.- Baker,R.J. and Nelder, J.A.(1978) The GLIM System ,Release 3,Generalized Linear Interactive Modelling,Numerical Algorithms Group, Oxford.

5.- Bennett, S. and Whitehead J.,(1983) Fitting Logistic and Log-Logistic Regression models o censored data using GLIM. GLIM Newsletter,4,12-19.

6.- Bennett S. (1983) Log-Logistic Regression Models for Survival Data. Appl.Statist.Vol 32,No. 2,165-171.

7.- Chiang, C.L.(1968) Introduction to stochastic Processes in Biostatistics, New York. John Wiley and Sons.

8.- Clayton, D.G.(1983) Fitting a General Family of Failure-Time Distributions using GLIM. Appl.Statist,Vol. 32,No. 2:102-109.

9.- Clayton,D. and Cuzick, J. (1985) The EM algorithm for Cox's Regression Model using GLIM. Appl. Statist,Vol. 34. No.2:148-156.

10.- Cox, D. R.(1972) Regression Models and Life Tables (with discussion). Journal of the Royal Statistical Society.Ser. B,34:187-220.

11.- Cox, D.R. and Oakes,D. (1984) Analysis of survival Data.London New York. Chapman and Hall.(University Press, Cambridge).

12.- Elandt-Jonhson, R.C. y Johnson, N.H.(1980) Survival Models and Data Analysis, J.Wiley and Sons, New York.

13.- Epstein,B. and Sobel, M.(1953) Life testing. Journal of the American Statistical Association 48:486-502.

14.- Gross, A.J. and Clark,V.A.(1975) Survival Distributions: Reliability Applications in the Biomedical Sciences, New York: John Wiley and Sons.

TESIS CON.
FALLAS DE ORIGEN

15.- Jordan, C.W.(1952) Life Contingencies, Chicago, The society of Actuaries.

16.- Kalbfleisch, J.D. and Prentice R.L.(1973) Marginal Likelihoods based on Cox's regression and life model, Biometrika,60,2,267.

17.- Kalbfleisch, J.D. and Prentice, R.L. (1980) The statistical Analysis of Failure Time Data, J. Wiley and Sons, New York.

18.- Laird, N. Olivier, D.(1981) Covariance analysis of censored Survival Data Using Log-Linear Analisis Techniques.Journal of the American Statistical Association.76:231-240.

19.- McCullagh, P. and Nelder, J. A.(1983) Generalized Linear Models. London New York. Chapman and Hall. (University Press ,Cambridge).

20.- Miller, R.G. (1981) Survival Analysis, J. Wiley and Sons, New York.

21.- Nelder, J.A. and Wedderburn, R.W.(1972) Generalized Linear Models. Journal of the Royal Statistical Society. Ser. A,135:370-384.

22.- Nelder, J.A.(1974) Log Linear Models for Contingency Tables: A Generalization of Classical Least Squares.Appl. Statist. Vol. 23,No. 3: 323.

23.- Roger, J.H. and Peacock, S.D.(1982) Fitting the scale as a GLIM parameter for weibull,extreme value,logistic and Log-Logistic regression models with censored data. Glim Newsletter 6,pag:30-37.

24.- Schoenfeld, D.(1980) "Chi-squared goodness of-fit test for the proportional hazard regression model. Biometrika (1980) 67,1,pag. 145-53.

25.- Taulbee, J.(1979) A General Model for Hazard Rate whit covaribles. Biometrics Junio, 35,pag:439-450.

26.- Wedderburn, R. W. M.(1974) Generalized Linear Models Specified in terms of contraints . Journal of the Royal Statistical Society. Ser. B,36,449-454.

27.- Whitehead J.(1980) Fitting Cox's Regression Model to Survival Data Using GLIM. Appl. Statist.,vol. 29, No. 3, 268-75.