Análisis de Sobrevivencia utilizando el lenguaje R[1]

Rafael Eduardo Borges Peña[2]

Simposio de Estadística 2005
Paipa, Boyacá, Colombia, 1 al 5 de Agosto, 2005

[1]Trabajo financiado a través del Consejo de Desarrollo Científico, Humanístico y Tecnológico de la Universidad de Los Andes (CDCHT-ULA) (Proyecto Código E-199-02-09-C)

[2]Profesor Asistente, Escuela de Estadística, Facultad de Ciencias Económicas y Sociales, Universidad de Los Andes, Mérida 5101, Venezuela. e-mail: borgesr@ula.ve

Prefacio

Este material ha sido elaborado para el cursillo titulado **"Análisis de sobrevivencia utilizando el lenguaje R"** que ha sido programado en el XV SIMPOSIO DE ESTADÍSTICA de la Universidad Nacional de Colombia.

El cursillo está diseñado para ser tomado por personas que no tienen conocimientos acerca del tópico estadístico conocido como análisis de sobrevivencia ni del lenguaje de procesamiento estadístico R, y ha sido estructurado de tal manera de dar una introducción tanto al análisis de supervivencia, como del lenguaje R, y exponer las principales herramientas para llevar un análisis de supervivencia mediante el uso del lenguaje R.

El presente material ha sido dividido en capítulos claramente diferenciadas uno del otro.

En el primer capítulo se presenta una breve introducción al Lenguaje R.

En el segundo capítulo se presentan los recursos disponibles en el Lenguaje R para llevar a cabo un análisis de sobrevivencia, haciendo énfasis en las principales funciones contenidas en la librería survival.

El tercer capítulo presenta una visión introductoria del análisis de

supervivencia en el cual se presenta mecanismos de censura y truncamiento, algunas definiciones básicas.

En el cuarto capítulo, se presentan métodos de estimación de la función de sobrevivencia a través del estimador de Kaplan y Meier y el estimador de Fleming y Harrington, comparación de dos funciones de sobrevivencia, sobrevida media, sobrevida mediana, acompañado de un ejemplo ilustrativo utilizando el Lenguaje R.

En el quinto capítulo se estudia el modelo de regresión de Cox acompañado de la verificación de sus supuestos y culmina con un ejemplo de la implementación de los tópicos a través del Lenguaje R.

El sexto capítulo introduce los modelos paramétricos, estimación de los parámetros, selección del mejor modelo, acompañado de un ejemplo.

Finalmente debemos enfatizar que este es un material que este cursillo pretende servir de motivación para que algunos de los asistentes decidan seguir el estudio de los métodos de a nálisis de sobrevivencia y del lenguaje R.

Índice General

Capítulo 1

Introducción al lenguaje R

El lenguaje R es un lenguaje de computación formal diseñado para ser utilizado en la manipulación y análisis de datos que posee una serie de facilidades gráficas. Es además un lenguaje de licencia gratuita.

El lenguaje R puede ser considerado como un programa orientado a objetos debido a que la filosofía del lenguaje es que el resultado de la evaluación de cada función genera un objeto, que posee a su vez una serie de atributos.

El lenguaje R puede ser utilizado de manera interactiva, pudiéndose obtener resultados con cada línea de comandos, esta característica la hace diferente de otros paquetes importantes como los son el sistema SAS y el SPSS en su versión de programación.

Es además, un lenguaje con mucho potencial para el desarrollo de dispositivos gráficos y posee además un buen nivel de manipulación de datos pero quizás en este aspecto no es tan poderoso como por ejemplo el SAS, lo cual no constituye un problema ya que la importación y exportación de datos desde y hacia otros sistemas está

completamente resuelta, también existe la posibilidad de comunicarse con otros lenguajes poderosos como lo son PERL y PYTHON, con los cuales las posibilidades de manejos con cualquier clase de estructuras de datos son prácticamente ilimitadas.

El Lenguaje R es una iniciativa de desarrollo escrito inicialmente en 1996 por Robert Gentleman y Ross Ihaka de la Universidad de Auckland de Nueva Zelandia que se basó en el ambiente del lenguaje S. El lenguaje S es un lenguaje ideado a finales de los ochenta por personas ligadas a los Laboratorios Bell (Chambers, Becker y otros) y que fue comercializado con el nombre de S-PLUS por la compañía de software de Seattle STATSCI. Posteriormente fue comercializado por MATHSOFT, por LUCENT TECNOLOGY y más recientemente por INSIGHTFUL CORPORATION.

Recientemente (18 de Abril de 2005), fue liberada la versión 2.1.0 del lenguaje R (R Development Core Team, 2005), esta versión tiene incorporado el sistema básico y más de 25 paquetes considerados como estándares y recomendados. El acceso a la base del lenguaje puede hacerse a través de la página principal del proyecto R (http://www.r-project.org) o a través de los servidores espejos (mirror sites) de la red Comprehensive R Archive Network (http://cran.r-project.org).

Otra ventaja del lenguaje R es la gran cantidad de paquetes contribuidos disponibles en la página de CRAN, actualmente hay disponibles más de 500. Un aspecto importante de resaltar es que cada paquete viene acompañado de su manual en formato pdf, el cual puede descargarse de la página de CRAN.

1.1 Instalación del lenguaje R.

El lenguaje R bajo Windows, se instala ejecutando el archivo de instalación (rw2010.exe) y siguiendo las instrucciones de instalación, este archivo se puede bajar de cualquiera de los servidores de CRAN, este archivo ejecutable tiene un tamaño de 25 Mb y ocupa un espacio en disco máximo de un poco ms de 50 Mb, en su instalación completa (full). Los requerimientos de equipo no son muy exigentes, puede ser instalado en equipos de la familia x86 o superiores y funcionan con los sistemas operativos Microsoft Windows superiores a las versiones 3.11.

1.2 Instalación de los paquetes adicionales.

Los paquetes contribuidos se instalan directamente desde el ambiente de trabajo del lenguaje R, utilizando el menú Packages. Esto puede hacerse de dos maneras:
i) Directamente de las páginas de CRAN, para lo cual hay que estar conectado a la internet.
ii) A través de los archivos ejecutables previamente bajados en formato zip, en esta opción no es necesario estar conectado a internet.

Un aspecto interesante es que los paquetes que han sido instalados previamente pueden ser actualizados directamente de la página de CRAN, esta facilidad solo se puede utilizar si se está conectado a internet.

1.3 Ayudas y documentación del R.

R es un lenguaje que ofrece varios niveles de ayuda y estas pueden activarse a través de la línea de comandos o a través del menu Help.

A través de la línea de comandos pueden utilizarse las instrucciones:

help("topico") para obtener ayuda acerca del tópico específico, el inconveniente de esta instrucción es que hay que colocar exactamente el nombre del tópico.

La instrucción help.search("tópico") es más flexible porque se realiza una búsqueda del tópico en todas los paquetes que han sido bajados a al R de la máquina.

Las dos opciones anteriores también están disponibles a través del menú Help (R functions (text)... y, Search help..., respectivamente.)

En el menu Help también están disponibles una excelente ayuda en formato HTML, en el cual se van incorporando las ayudas de los paquetes que se van incorporando.

Otra forma de ayuda son los manuales en formato pdf, teniendo disponible a través del menú Help los relatvos a la base del R. Los manuales del resto de los paquetes pueden accesarse mediante el acrobat reader.

Existe también una serie de documentación no oficial de R que está disponible a través de la página de CRAN. Esta documentación es considerada como contribuida y hay material en diversos idiomas, incluyendo al castellano.

1.4 Acceso a datos internos disponibles.

Una gran parte de los paquetes tiene disponible una serie de datos que pueden ser trabajados, sobre todo en la etapa de aprendizaje del lenguaje R. La verificación de los datos disponibles en todos los

paquetes disponibles en el R que se está trabajando puede verse mediante la instrucción data() y para ver los datos de un paquete en específico debe escribirse la opción data("paquete").

1.5 Acceso a datos externos disponibles.

La forma más sencilla de acceder a datos externos es mediante la función read.table para lo cual se recomienda tener los datos en un archivo de texto delimitado por algún carácter, por defecto el espacio. Se recomienda además tener los nombres de las variables en la primera fila en cuyo caso hay que colocar la opción header=TRUE.

También existen facilidades para importar (y exportar) datos desde (y hacia) otros sistemas, una opción interesante es la del paquete foreign, que permite importar y exportar datos de Minitab, S, SAS, SPSS y Stata, entre otros.

1.6 La opción de asignación.

Todo comando que se ejecuta en R produce un resultado y para poder tener disponible este resultado debe hacerse un proceso de asignación, esto de hace mediante los caracteres menor que ($<$) y el guión ($-$), generando el efecto visual de una flecha, $<-$, el proceso de asignación se hace asignando el valor o el objeto a la derecha de $<-$ al objeto cuyo nombre es colocado a la izquierda de $<-$. El nombre del objeto al cual se le hace la asignación puede incluir cualquier carácter alfanumérico, incluyendo puntos y se recomienda que comience con un letra. El formato de la asignación es:

Nombre.del.objeto$<-$(función u objeto generado)

1.7 Verificación de objetos disponibles.

Los objetos disponibles pueden ser verificados mediante la instrucción objects().

1.8 Eliminación de objetos no deseados.

Cada proceso de asignación va generando un nuevo objeto, en caso de no nesecitar más un objeto este debe borrarse utilizando la instrucción rm("Nombre objeto a eliminar"). En el caso de que se quieran eliminar varios objetos, pueden colocarse toso ellos dentro del paréntesis separados por comas.

1.9 R diferencia las mayúsculas de las minúsculas.

Un aspecto que debe considerarse cuando se trabaja en R es que el lenguaje diferencia entre las mayúsculas y las minúsculas, por loe cada comando u objeto debe se escrito de manera exacta.

1.10 Datos faltantes en R.

Los datos faltantes para variables numéricas en R se suele especificar con el valor NA. No todos las funciones admiten la presencia de datos faltantes por lo que hay que revisar primero la documentación disponible para la función o efectuar pruebas correspondientes.

1.11 Comentarios en R.

El lenguaje R admite comentarios. Un comentario comienza con el carácter numeral (#), considerándose como comentario a todo lo

que aparezca el la línea de comando a la derecha de #.

1.12 Creación de datos en R.

El lenguaje R permite la creación de diversas estructuras de datos para el caso de que se tengan pocas variables y pocos individuos el proceso de creación puede hacerse a través de una lista de combinaciones por columna, mediante la instrucción:

list(cbind(var1=c(valor1,...,valorn),...,vark=c(valor1,...,valorn)))

Otra opción es la creación de vectores por separado y unirlos por columnas mediante la función cbind y luego eliminar los objetos de las variables (vectores) pero este procedimiento es menos eficiente que anterior.

Otra forma de crear datos es crearlos con un software externo, por ejemplo Microsoft excel, guardarlos como archivos de texto delimitados, teniendo cuidado de que el separador de enteros y decimales debe ser un punto, luego copiarlos en el sudirectorio del R y leerlos mediante la función read.table.

1.13 Carga y descarga de objetos.

Algunas funciones de R, necesitan tener accesible el objeto que contienes las variables a analizar, la carga en el ambiente se hace mediante la instrucción attach(objeto) y la descarga mediante la instrucción dettach(objeto).

1.14 Envío de gráficos a otros programas.

El envío de gráficos a otros programas se hace colocando el gráfico en el pisapapeles. Existen dos opciones para esto, colocándolo como metafile o como bitmap, el ícono de la cámara lo coloca como metafile.

1.15 Salida del lenguaje R.

Para salir del lenguaje R, se debe escribir la instrucción `q()`, seleccionar la opción Exit del menú File o hacer clic en la x colocada en el extremo superior derecho de la ventana. En este momento se pregunta si se desea guardar el espacio de trabajo, en caso de seleccionar si se graban de manera definitiva los objetos y los comandos que se han generado durante la sesión y si se selecciona no, se pierde la información de la sesión por lo que debe estar muy atento para tomar la decisión al momento de cerrar la sesión de R.

Capítulo 2

Análisis de sobrevivencia utilizando el lenguaje R

El análisis de sobrevivencia en el lenguaje R puede hacerse a través de un conjunto de paquetes especializados que se detallan a continuación:

bayesSurv: Modelos de Regresión Bayesianos.

cmprsk: Análisis de riesgos en competencia por subdistribuciones.

dblcens: Calcula estimadores máximo verosímiles no paramétricos para datos con doble censura.

eha: Sus siglas significan *event history analysis* y contiene funciones que permiten ajustar modelos de regresión en análisis de sobrevivencia.

emplink: Contiene test de razón de verosimilitudes empíricos para datos censurados y truncados.

Icens: Calcula estimadores máximo verosímiles no paramétricos para datos censurados y truncados.

intcox: Contiene un algoritmo convexo iterado de aminoramiento para datos con censura por intervalos.

kinship: Contiene funciones para modelos de Cox de efectos mixtos.

KMsurv: contiene los datos del libro de Klein y Moeschberger (1997).

msm: Que trabaja con modelos de Markov de múltiple estados continuos en el tiempo y que son útiles para algunos modelos de sobrevivencia multivariados.

muhaz: Contiene funciones que permiten hacer estimaciones de la función de riesgos.

relsurv: Contiene funciones que permiten ajustar modelos de regresión relativos en análisis de sobrevivencia.

smoothSurv: Trabaja con modelos de regresión con distribuciones de errores suavizadas.

survBayes: Permite ajustar modelos de riesgos proporcionales bajo un enfoque Bayesiano.

survival: Es el principal paquete para realiza Análisis de Sobrevivencia.

survnnet: Dedicado a Análisis de Sobrevivencia a través de redes neuronales.

survrec: Contiene funciones que permiten estimar funciones de sobrevida para datos de eventos recurrentes.

zicount Contiene funciones que permiten ajustar modelos de regresin para datos de conteo censurados.

Existen otros paquetes que aunque no están orientados exclusivamente al Análisis de Sobrevivencia, contienen funciones útiles, uno de ellos es de miscelaneas de Harrel (**Hmisc**).

De los paquetes mencionados anteriormente, el más utilizado, que también tiene el estatus de recomendado, es el **survival**, cuya versión 2.17 está disponible desde el 6 de Abril del presente año. El paquete survival es una librería desarrollada por Thomas Lumley a partir del código para S desarrollado inicialmente por Terry Therneau (S original by Terry Therneau and ported by Thomas Lumley, 2005), puede verse también el texto de Therneau y Grambsch (2000).

2.1 El paquete survival

El paquete survival permite llevar a cabo análisis de sobrevivencia para datos que presentan diversos mecanismos de censura. Este es un paquete que tiene la característica de ser un paquete con el estatus de recomendado el cual ya viene incorporado en la versión 2.1.0 de R (versión 2.17), sin embargo, es recomendable estar atentos para actualizarlo de manera permanente. Algunas de las funciones de este serán descritas brevemente en la subsecciones siguientes. Para ejecutar cualquiera de las funciones de este paquete es necesario invocar la librería mediante la instrucción:

```
library(survival)
```

2.1.1 La función Surv

La función Surv permite crear objetos tipo survival, la estructura para datos que presentan censura por la derecha mediante:

```
Surv(time, event)
```

En la cual time representa el tiempo y event representa el estatus de censura, considerado como cero (0) para datos censurados y como uno (1) cuando el evento es observado.

Una estructura más completa de la función Surv, que es útil para otros tipos de censuras es:

```
Surv(time, time2, event, type=, origin=0)
```

En donde time representa el tiempo de inicio de la observación, time2 el tiempo de finalización, se asume que los intervalos de tiempos son abiertos en su extremo inferior y cerrados en su tiempo superior, es decir (time, time2], event es la condición de ocurrencia del evento que depende del tipo de censura (type), que por defecto es censura por la derecha, y origin es una utilidad que permite trabajar bajo el enfoque de los procesos de conteo.

2.1.2 La función survfit

La función survfit permite obtener estimación de la función de sobrevivencias utilizando el método de Kaplan y Meier (opción por defecto) o de Fleming y Harrington. También permite predecir la función de sobrevivencia para modelos de Cox. La estructura de la función survfit es:

```
survfit(formula, data, weights, subset, na.action,newdata,
individual=F,conf.int=.95, se.fit=T, type=c("kaplan-meier",
```

"fleming-harrington", "fh2"), error=c("greenwood","tsiatis"),
conf.type=c("log","log-log","plain","none"),
conf.lower=c("usual", "peto", "modified"))

la cual posee una serie de opciones que pueden ser revisadas en el manual (S original by Terry Therneau and ported by Thomas Lumley, 2005) o en la ayuda (help("survfit"))

Con la función survfit puede obtenerse diversa información:

Con print(survfit(. . .)) o directamente con survfit(. . .) se obtienen las medidas resumen.

Con summary(survfit(. . .)) se obtiene la función de sobrevida estimada.
Con plot(survfit(. . .)) se obtiene el gráfico de la función de sobrevivencia estimada. En esta función pueden controlarse un serie de opciones gráficas, se recomienda ver la ayuda correspondiente para más detalles.

Con names(survfit(. . .)) se obtiene el nombre de cada uno de los atributos de la función survfit. Esta función es útil para seleccionar atributos por separados o para realizar cálculos posteriores cuando sea necesario.

2.1.3 La función survdiff

La función survdiff permite efectuar contrastes de hipótesis para verificar la igualdad o diferencia de dos o más curvas de sobrevivencias, basados en las familias de pruebas G-rho propuestas por Harrington y Fleming (1982). La estructura de la función survdiff es:

survdiff(formula, data, subset, na.action, rho=0)

Para más detalles ver la ayuda correspondiente (help("survdiff")).

2.1.4 La función coxph

La función coxph permite ajustar modelos de regresión de Cox. Permite también ajustar modelos con variables dependientes del tiempo, modelos estratificados, modelos de múltiples eventos por individuo y otras extensiones derivadas del enfoque basado en los procesos de conteo. La estructura de la función coxph es:

coxph(formula, data=parent.frame(), weights, subset,
na.action, init, control, method=c("efron","breslow","exact"),
singular.ok=TRUE, robust=FALSE,
model=FALSE, x=FALSE, y=TRUE,...)

Para mayores detalles puede consultar la ayuda correspondiente.

La función coxph puede combinarse con otras funciones que permiten obtener la siguiente información:

Con print(coxph(. . .)) o directamente con coxph(. . .) se obtienen las los contrastes para verificar la adecuacidad del modelo de Cox ajustado.

Con summary(coxph(. . .)) se obtiene un poco más de detalles de los contrastes.

Con summary(survfit(coxph(. . .))) se obtiene la función de sobrevivencia ajustada por el modelo de Cox.

Con plot(survfit(coxph(. . .))) se obtiene el gráfico de la función de

sobrevida ajustada por el modelo de Cox. En esta función pueden controlarse un serie de opciones gráficas, se recomienda ver la ayuda correspondiente para más detalles.

Con names(coxph(. . .)) se obtiene el nombre de cada uno de los atributos de la función coxph. Esta función es útil para seleccionar atributos por separados o para realizar cálculos posteriores cuando sea necesario.

Otra funciones importantes que funcionan con la función coxph son lo son las funciones cox.zph y la función residuals (o resid), las cuales serán explicadas en la próximas dos subsecciones.

2.1.5 La función cox.zph

La función cox.zph permite llevar a cabo el contraste de hipótesis de riesgos proporcionales, las salidas directas presentan el contraste global y de cada una de las covariables en el modelo. La hipótesis nula es el cumplimiento del supuesto de riesgos proporcionales, asociado a que los betas son ceros. La estructura de esta función es:

```
cox.zph(fit, transform="km", global=TRUE)
```

Esta función puede combinarse con la función plot para obtener la distribución de los betas, para lo cual se utiliza la siguiente estructura:

```
plot(x, resid=TRUE, se=TRUE, df=4, nsmo=40, var, ...)
```

donde x es un objeto de tipo cox.zph y var permite identificar la covariable que se va a representar de forma gráfica, en S-PLUS no es necesario hacer esta declaración porque se genera un gráfico compuesto donde se representan los gráficos para cada una de las

covariables, consulte la ayuda para detalles adicionales. Para entender un poco más como funciona este comando puede ver el ejemplo del presente curso.

2.1.6 La función residuals

Otra función importante asociada a los objetos del tipo coxph es la función residuals, o en su formato más corto resid. Esta función permite calcular los residuos de martingala, de puntajes (score), de tipo desvío (deviance) y de Schoenfeld. La estructura de esta función es:

residuals(object, type=c("martingale", "deviance", "score", "schoenfeld", "dfbeta", "dfbetas", "scaledsch","partial"), collapse=FALSE, weighted=FALSE, ...)

donde object es un objeto de tipo coxph.

Para mayores detalles consulte la ayuda correspondiente y para ver la implementación puede verse el ejemplo de este curso.

2.1.7 La función survreg

La función survreg permite ajustar modelos de regresión paramétricos utilizados en análisis de sobrevivencia y confiabilidad. La estructura de la función survreg es:

survreg(formula=formula(data), data=parent.frame(), weights,
subset,na.action,dist="weibull", init=NULL, scale=0,
control=survreg.control(),parms=NULL,model=FALSE, x=FALSE,
y=TRUE, robust=FALSE, ...)

la cual posee una serie de opciones que pueden ser revisadas en el manual (S original by Terry Therneau and ported by Thomas

Lumley, 2004) o en la ayuda (help("survreg"))

Las distribuciones que se pueden modelar directamente a través de la función survreg son la Weibull, la exponencial, la gaussiana o normal, la lognormal, la logística, la loglogística.

Con la función survreg puede obtenerse diversa información:

Con print(survreg(. . .)) o directamente con survreg(. . .) se obtiene una información bastante completa del ajuste.

Con names(survreg(. . .)) se obtiene el nombre de cada uno de los atributos de la función survreg. Esta función es útil para seleccionar atributos por separados o para realizar cálculos posteriores cuando sea necesario.

Con summary(survreg(. . .)) se obtiene información general acerca de los atributos del objeto tipo survreg.

2.1.8 La función survreg.distributions

Es una función que permite declarar otros modelos paraméricos, principalmente los pertenecientes a la familia de localización y escala.

Capítulo 3

Introducción al análisis de sobrevivencia.

El análisis de sobrevivencia consiste en un conjunto de técnicas para analizar el tiempo de seguimiento hasta la ocurrencia de un evento de interés. Este tiempo de seguimiento hasta que ocurra el evento de interés, también denominado tiempo de vida puede observarse completa o parcialmente. Un caso poco frecuente en la práctica es aquel en que se observan los individuos desde un evento inicial hasta el evento de final o de ocurrencia del fenómeno que se desea observar. A la ocurrencia del evento de interés se le suele denominar falla o muerte.

Ahora bien, es posible, y muy frecuente en la práctica encontrarse con situaciones en que se cuenten con observaciones incompletas de los períodos que transcurren entre el tiempo inicial y el tiempo final. Esto puede darse por censura o por truncamiento, y es precisamente bajo la presencia de censura o truncamiento que el análisis de sobrevivencia cobra una importancia primordial.

Debido a la presencia de censura y/o truncamiento al análisis de

supervivencia se le conoce también como análisis de datos censurados y/o truncados. Otro de los nombre que recibe este análisis es análisis de confiabilidad (reliability), derivados de estudios de distribuciones del tiempo de vida en procesos industriales o de ingeniería. También puede ser denominado como análisis del tiempo hasta la ocurrencia de un evento.

En este Capítulo se presenta una visión introductoria del análisis de supervivencia, incluyéndose aspectos relacionados con censura y truncamiento, seguido de una serie de definiciones básicas.

En capítulos posteriores se presentará el estimador de Kaplan y Meier, métodos de comparación de funciones de riesgos, el modelo de regresión de Cox y modelos paramétricos.

3.1 Censura y truncamiento

Existen dos mecanismos que no hacen posible la observación completa de los tiempos de seguimiento, como lo son la censura y el truncamiento. En cuanto a la censura existen dos tipos: censura tipo I en la cual los individuos son observados hasta un tiempo determinado y, la censura tipo II en la cual los individuos son observados hasta que ocurran un nmero determinado de fallas o eventos de interés. Los mecanismos de censura (tipo I) y truncamiento más frecuentes son presentados a continuación:

i) Censura por la derecha: Se presenta cuando hasta la ltima observación que se le hace al individuo, an no se ha ocurrido el evento que se desea observar. Existen varias razones para que se presente este tipo de censura:
- Que hasta el momento de la finalización del estudio no haya ocurrido el evento, esto ocurriría en el caso de que el período de seguimiento sea finito.

- Que el individuo haya abandonado el estudio.
- Que haya ocurrido en el individuo otro evento que imposibilite la ocurrencia del evento que se desea observar.

ii) Censura por la izquierda: Es poco comn en análisis de supervivencia, se presenta cuando para la primera observación que se realiza sobre el individuo ya ha ocurrido el evento que se desea observar. Este tipo de censura suele confundirse con el truncamiento por la izquierda o la entrada tardía.

iii) Censura por intervalos: Se presenta cuando solo se sabe que al individuo le ocurre el evento de interés entre un instante t_i y un tiempo t_j.

iv) Entrada tardía al estudio (truncamiento por la izquierda): Se presenta cuando el individuo comienza a observarse posteriormente al verdadero evento inicial.

vii) Truncamiento por la derecha: Se presenta cuando sólo se incluyen los individuos que presentan el evento o falla de interés.

Para obtener un panorama general de los distintos tipos de censura puede verse el libro de Andersen y colaboradores (Andersen et al., 1993) o el de Klein y Moeschberger (1997).

3.2 Definiciones básicas

3.2.1 Función de supervivencia

La función de supervivencia se define como la probabilidad de que una persona sobreviva (no le ocurra el evento de interés) al menos hasta el tiempo t.

Una definición más formal puede darse de la siguiente manera: sea T una variable aleatoria positiva (o no negativa) con función de distribución $F(t)$ y función de densidad de probabilidad $f(t)$.

La función de supervivencia $S(t)$ puede escribirse como:

$$S(t) = 1 - F(t) = P[T > t]$$

3.2.2 Funciones de riesgos (hazard)

La función de razón de riesgos o tasa instantánea de fallas$\lambda(t)$ se define como el cociente entre la función de densidad y la función de supervivencia:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

Se interpreta como la probabilidad de que a un individuo le ocurra el evento de interés en la siguiente unidad de tiempo Δt dado que ha sobrevivido hasta el tiempo t.

Dicha función proviene de la tasa media de fallas, como sigue:

Dada la Probabilidad condicional de fallas en el período $(t; t + \Delta t)$, dado que la persona sobrevive en el período $(0; t)$, la tasa media de fallas (TMF) se define como:

$$\text{TMF} = \frac{F(t+\Delta t)-F(t)}{\Delta t}\frac{1}{S(t)}$$

Tomando límites para $\Delta t \to 0$, queda:

$$\lambda(t) = \lim_{\Delta t \to 0} \text{TMF} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}$$

De la expresión anterior, se puede obtener la funcin de sobrevida, mediante:

$$S(t) = \frac{f(t)}{\lambda(t)}$$

La función de riesgo acumulada $\Lambda(t)$ se define como:

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$$

De la expresión anterior, puede obtenerse, x puede obtenerse la función de sobrevida, a partir de la función de riesgo o de la función de riesgo acumulada, mediante la siguiente fórmula:

$$S(t) = e^{-\int_0^t \lambda(t)dt} = e^{-\Lambda(t)}$$

Como habíamos planteado anteriormente, lo que distingue al análisis de sobrevivencia es la presencia de la censura. El caso más comn de censura es la censura por la derecha, que se caracteriza porque sólo se sabe que el tiempo de ocurrencia del evento de interés es mayor que el ltimo tiempo de observación.

Los datos de supervivencia suelen presentarse en la forma (t_i, δ_i) donde t_i es el tiempo de observación y, $\delta_i = 0$ si la observación es censurada y $\delta_i = 1$ cuando se observa la ocurrencia del evento de interés.

Capítulo 4

Estimación de la función de sobrevivencia

En este capítulo se presentan dos estimadores para la función de sobrevida: el de Kaplan y Meier y el de Fleming y Harrington. Posteriormente, se presentan mtodos para comparar funciones de sobrevida y se introducen los conceptos de sobrevida media y sobrevida mediana.

4.1 Estimador de Kaplan y Meier

La presencia de datos censurados o truncados hace que la función de sobrevivencia no pueda ser obtenida directamente a través de argumentos probabilísticos haciéndose necesario el uso de algunos estimadores. Existen varias formas de estimar la función de sobrevivencia, entre los más conocidos son los basados en tablas de vida, entre el que se incluye el estimador actuarial y el estimador de Kaplan y Meier, que es más práctico, porque no es necesario trabajar con períodos de tiempos, sino que los mismos tiempos de observación van contribuyendo a la estimación de la función de supervivencia.

El estimador de Kaplan y Meier (1958) es el estimador de la función de sobrevivencia más utilizado y se define para el caso en que los datos puedan presentar censura por la derecha como:

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \frac{r(t_i) - d(t_i)}{r(t_i)}$$

donde $r(t_i)$ y $d(t_i)$ son el número de individuos en riesgo y el número de muertes (o de ocurrencia del evento de interés) en el momento t_i.

La varianza del estimador de Kaplan y Meier se obtiene a través de la fórmula de Greenwood (1926):

$$V\left(\hat{S}_{KM}(t)\right) = \hat{S}^2_{KM}(t) \sum_{t_i \leq t} \frac{d(t_i)}{r(t_i)[r(t_i) - d(t_i)]}$$

El intervalo de confianza del 95% de escala plana (o de identidad), llamado así porque es obtenido de manera estándar al que se obtiene cualquiera de los intervalos de confianza, sin utilizar ninguna transformación, se obtiene mediante:

$$\hat{S}_{KM}(t) \pm 1.96 \text{ee}\left(\hat{S}_{KM}(t)\right)$$

donde $\text{ee}\left(\hat{S}_{KM}(t)\right)$ es el error estándar de estimación del estimador de Kaplan y Meier.

4.2 Estimador de Fleming y Harrington

Un estimador de la función de la función de sobrevida puede obtenerse a partir del estimador de Nelson y Aalen.

El estimador Nelson y Aalen (Nelson, 1969) es una estimador de la función de riesgo acumulado que puede calcularse mediante:

$$\hat{\Lambda}_N(t) = \sum_{t_i \leq t} \frac{d(t_i)}{r(t_i)}$$

y bajo el enfoque de procesos de conteo, toma la forma:

$$\hat{\Lambda}_N(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{r(s)}$$

De donde puede hallarse el estimador de función de sobrevida de Fleming y Harrington (Fleming y Harrington, 1984 y 1991), mediante:

$$\hat{S}_{FH}(t_j) = e^{-\hat{\Lambda}_N(t_j)}$$

4.3 Comparación de las funciones de sobrevivencia

La comparación de dos curvas de sobrevida se efectúa a través de contrastes basados en tablas de contingencia como la siguiente:

Tabla No. 1. Tabla usada para el contraste de igualdad de funciones de sobrevivencia en dos grupos en el tiempo de observación t_i

	Grupo		
Evento	1	0	Total
Muerte	$d_1(t_i)$	$d_0(t_i)$	$d(t_i)$
No muerte	$r_1(t_i) - d_1(t_i)$	$r_0(t_i) - d_0(t_i)$	$r(t_i) - d(t_i)$
En riesgo	$r_1(t_i)$	$r_0(t_i)$	$r(t_i)$

Donde por comodidad se han definido los grupos, como 1 y 0, correspondiendo estos grupos a cada una de las dos curvas de supervivencia.

Para construir el estadístico de contraste basta con calcular el número

esperado de muertes y la varianza estimada del número de muertes para uno de los grupos; por ejemplo, para el grupo 1 el número esperado de muertes se calcula de la siguiente manera:

$$\hat{e}_1\left(t_i\right) = \frac{r_1(t_i)d(t_i)}{r(t_i)}$$

La varianza estimada de $d_i\left(t_i\right)$ está basada en la distribución hipergeométrica y para el grupo 1 está definida como:

$$\hat{V}\left(d_1\left(t_i\right)\right) = \frac{r_1(t_i)r_0(t_i)(r(t_i)-d(t_i))}{r^2(t_i)(r(t_i)-1)}$$

Finalmente, el estadístico de contraste se define de la siguiente manera:

$$Q = \frac{\left[\sum_{i=1}^{m} w_i(d_1(t_i)-\hat{e}_1(t_i))\right]^2}{\sum_{i=1}^{m} w_i^2\hat{V}(d_1(t_i))}$$

Puede demostrarse que el estadístico anterior se puede aproximar mediante una Chi cuadrado de un grado de libertad si el número de ocurrencias de eventos es grande.

Bajo la hipótesis nula que asume que las dos funciones de sobrevivencia son iguales. En esta fórmula m es el número de tiempos de ocurrencia de eventos en ambos grupos y w_i denota los pesos, que toman valores distintos dependiendo del test utilizado. En este curso sólo utilizaremos dos de los casos: el test de Mantel y Haenzel, mas conocido como el test de los rangos de logaritmos (log-rank test) y el test de Peto y Peto. Para una enumeración muy completa de los distintos test, basados en procesos de conteo (Andersen et al, 1993, Fleming y Harrington, 1991).

Test de Mantel y Haenzel: Como establecimos anteriormente, el más común de los test es de Mantel y Haenzel (o log-rank). Este test está diseñado para verificar igualdad o diferencia en la función

de sobrevivencia en todos los tiempos. En este test los pesos son iguales a 1, es decir, $w_i = 1$ (Mantel, 1966).

Test de Peto y Peto: Otro de los test comúnmente utilizados es el de Peto y Peto (Peto y Peto, 1972). Este test permite verificar igualdad o diferencia de las funciones de sobrevivencia en los tiempos iniciales. En este test los pesos toman la forma:

$$w_i = \tilde{S}\left(t_{i-1}\right) \frac{r(t_i)}{r(t_i)-1}$$

donde $\tilde{S}\left(t\right)$ es el estimador de la función de sobrevivencia definida por:

$$\tilde{S}\left(t\right) = \prod_{t_i \leq t} \left(\frac{r(t_i)+1-d(t_i)}{r(t_i)+1}\right)$$

Familia de tests G-rho de Fleming y Harrington: Otra forma de estudiar los test anteriores fue propuesta por Harrington y Fleming (Harrington y Fleming, 1982 y Fleming y Harrington, 1991). Esos dos autores sugieren pesos de la forma:

$$w_1 = \left[\hat{S}_{KM}\left(t_{i-1}\right)\right]^{\rho}$$

y haciendo $\rho = 0$ se tiene que $w_i = 1$ (test log-rank) y, si $\rho = 1$, se obtiene el test de Peto y Peto. Esta manera de definir los pesos es la forma como trabaja el lenguaje R.

4.4 Sobrevida media y mediana

4.4.1 Sobrevida media

La sobrevida media o media de la supervivencia puede ser estimada mediante la siguiente expresión:

$$\hat{\mu} = \int_0^T \hat{S}_{KM}\left(t\right) dt$$

donde T es tiempo máximo de seguimiento observado durante el estudio.

La varianza de la media es:

$$\text{var}\left(\hat{\mu}\right) = \int_0^T \left(\int_0^T \hat{S}_{KM}\left(u\right) du \right)^2 \frac{dN(t)}{r(t)(r(t)-N(t))}$$

donde $N\left(t\right) = \sum N_i\left(t\right) = d\left(t\right)$ es el número total de muertes (o de ocurrencia del evento de interés hasta el tiempo t) y $r\left(t\right)$ es el número de individuos en riesgo en el tiempo t.

4.4.2 Sobrevida mediana

La sobrevida mediana o mediana de la supervivencia se define como el primer tiempo t que satisface la siguiente condición:

$$\hat{S}_{KM}\left(t\right) \leq 0.5$$

4.5 Ejemplo 4.1

En esta seccin se presenta la primera parte de un de un análisis de sobrevivencia utilizando el lenguaje R, para ello se ha trabajado con una versión reducida de los datos correspondientes a los pacientes de diálisis peritoneal analizados en la tesis de maestría de Borges (2002), que se encuentran en el archivo dpa.txt.

El análisis que se presenta en este capítulo corresponde al estimador de Kaplan y Meier y comparación de funciones de sobrevivencia.

4.5.1 Estructura del archivo de datos dpa.txt

El archivo dpa.txt contiene la información de 246 pacientes que acudieron al Servicio de diálisis peritoneal del Hospital Clínico Uni-

versitario de Caracas entre los años 1980 y 2000. La variables seleccionadas para este archivo fueron:

Variable	Descripción
orden:	Orden de los individuos en la base de datos.
sexofm:	Sexo (0 corresponde al sexo femenino y 1 al sexo masculino)
diabetes:	Diabetes mellitus (1 corresponde a un paciente diabético y 0 a uno no diabético)
meses:	Meses de seguimiento en diálisis peritoneal.
censor2:	Condición de Censura (1 denota la muerte y 0 denota los datos censurados)
edad:	Edad del paciente al comienzo de la diálisis).
quetelet:	índice de Quetelet.

4.5.2 Lectura de los datos en R

La lectura de los datos y su correspondiente asignación al objeto dpa se hace mediante la instrucción:

```
> dpa< −read.table("dpa.txt",header=TRUE)
```

Posteriormente y antes de comenzar a ejecutar las funciones del paquete survival debe ejecutarse el comando:

```
> library(survival)
```

4.5.3 Estimación de la función de sobrevivencia a través del estimador de Kaplan y Meier

La obtención del objeto que contiene la información de la estimación de la función de sobrevivencia a través del método de Kaplan y Meier se hace, luego de cargar el objeto dpa, mediante la instruccin:

```
> attach(dpa)
```

Para luego obtener el estimador de Kaplan y Meier mediante el comando:

```
> km1< −survfit(Surv(meses,censor2))
```

Para visualizar el resumen de la estimación debe escribirse la instrucción:

```
> print(km1)
```

o simplemente mediante:

```
> km1
```

obteniéndose la siguiente salida:

Call: survfit(formula = Surv(meses, censor2))

n	events	median	0.95LCL	0.95UCL
246	64	61	55	Inf

Los nombres del objeto km1 se obtienen mediante el comando:

```
> km1
```

obteniéndose el siguiente resultado:

```
[1] "n" "time" "n.risk" "n.event" "surv" "type" "std.err"
[8] "upper" "lower" "conf.type" "conf.int" "call"
```

La estimación de la función de supervivencia se obtiene mediante la instrucción:

> summary(km1)

Obteniéndose la siguiente salida:

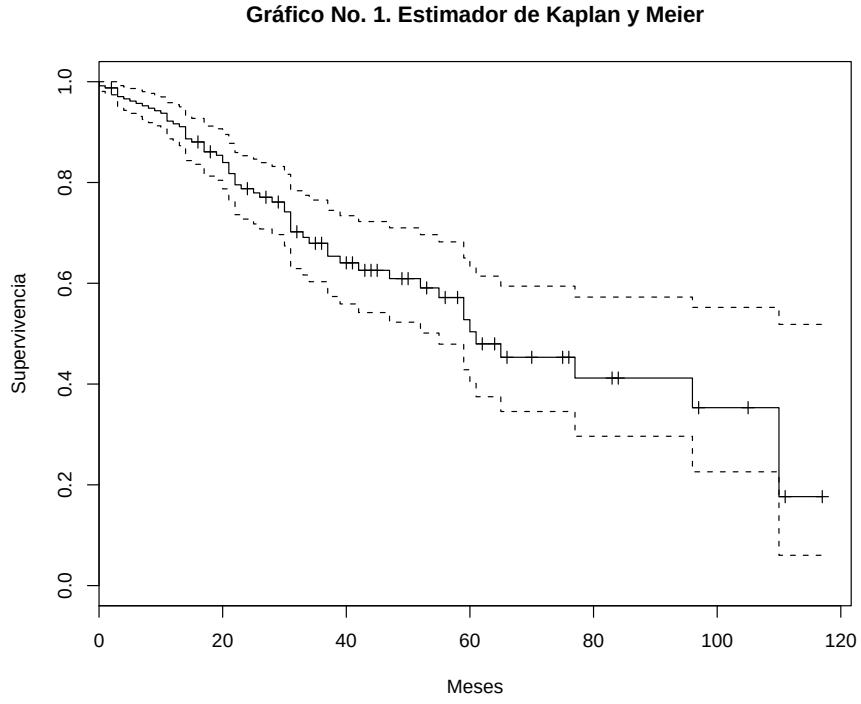
Call: survfit(formula = Surv(meses, censor2))

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	246	2	0.992	0.00573	0.9807	1.000
1	240	1	0.988	0.00704	0.9740	1.000
3	228	4	0.970	0.01102	0.9490	0.992
4	221	1	0.966	0.01182	0.9431	0.989
5	215	1	0.962	0.01259	0.9372	0.987
6	209	1	0.957	0.01334	0.9311	0.983
7	202	1	0.952	0.01409	0.9250	0.980
8	197	1	0.947	0.01483	0.9187	0.977
9	193	1	0.942	0.01554	0.9125	0.973
10	188	1	0.937	0.01625	0.9061	0.970
11	180	3	0.922	0.01831	0.8866	0.958
12	171	1	0.916	0.01898	0.8800	0.954
13	161	1	0.911	0.01970	0.8729	0.950
14	151	4	0.887	0.02257	0.8435	0.932
15	144	1	0.880	0.02324	0.8361	0.927
17	135	3	0.861	0.02532	0.8127	0.912
19	124	1	0.854	0.02605	0.8044	0.907
20	119	2	0.840	0.02752	0.7873	0.895
21	115	3	0.818	0.02956	0.7617	0.878
22	110	3	0.795	0.03143	0.7361	0.859
23	104	1	0.788	0.03205	0.7274	0.853
25	94	1	0.779	0.03278	0.7177	0.846
26	90	1	0.771	0.03354	0.7077	0.839
28	81	1	0.761	0.03445	0.6966	0.832
30	78	2	0.742	0.03623	0.6739	0.816
31	75	4	0.702	0.03933	0.6291	0.784
33	63	1	0.691	0.04025	0.6164	0.775
34	59	1	0.679	0.04124	0.6031	0.765
37	53	2	0.654	0.04348	0.5737	0.745
39	50	1	0.641	0.04453	0.5589	0.734
42	43	1	0.626	0.04592	0.5418	0.722
47	38	1	0.609	0.04757	0.5227	0.710
52	33	1	0.591	0.04958	0.5011	0.696
55	31	1	0.572	0.05152	0.4791	0.682
59	26	2	0.528	0.05616	0.4283	0.650
60	22	1	0.504	0.05850	0.4012	0.632
61	21	1	0.480	0.06044	0.3748	0.614
65	18	1	0.453	0.06268	0.3455	0.594
77	11	1	0.412	0.06920	0.2963	0.573
96	7	1	0.353	0.08054	0.2258	0.552
110	4	2	0.177	0.09701	0.0601	0.518

Finalmente, el gráfico de la función de supervivencia puede construirse mediante el comando:

> plot(km1,xlab="Meses",ylab="Supervivencia", main="Gráfico No. 1. Estimador de Kaplan y Meier")

Obteniéndose el siguiente gráfico:



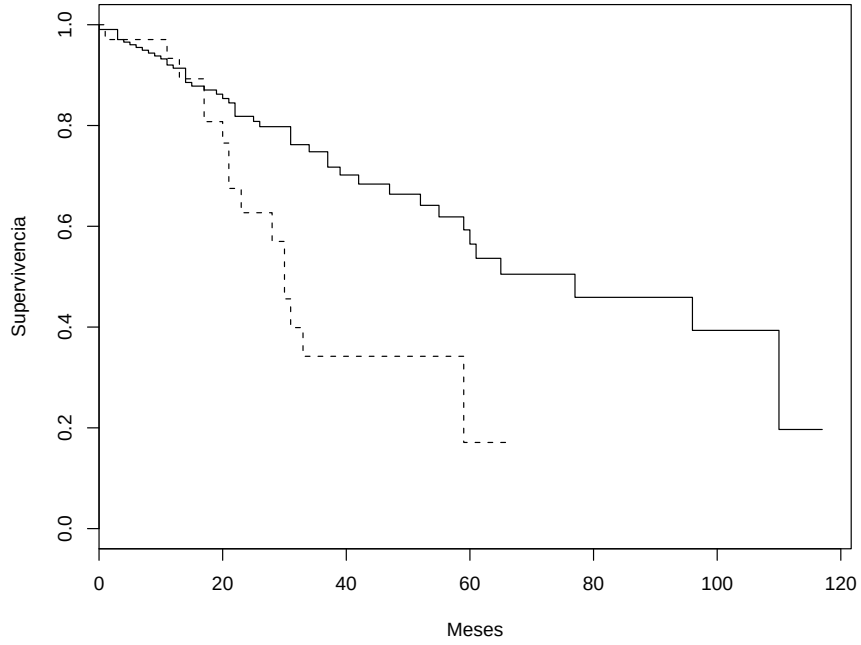
4.5.4 Comparación de funciones de supervivencia

Suponga que queremos comparar las funciones de supervivencia de los pacientes diabéticos y los no diabéticos, para ello construyamos un gráfico donde se observe las estimaciones de Kaplan y Meier para los pacientes diabéticos y no diabéticos, esto lo haremos mediante las instrucciones:

```
> km2< −survfit(Surv(meses,censor2)∼diabetes)
> plot(km2,xlab="Meses",ylab="Supervivencia", main="Gráfico
No. 2. Estimador de Kaplan y Meier \n para pacientes diabéticos y no
diabéticos",lty=c(1,2),mark.time=FALSE)
> legend(75,0.9,legend=c("No diabéticos","Diabéticos"),lty=c(1,2))
```

Obteniéndose el gráfico:

Gráfico No. 2. Estimador de Kaplan y Meier para pacientes diabéticos y no diabéticos



Donde de observa que aparentemente ambas funciones de supervivencia son distintas.

Para comparar ambas funciones de supervivencia se debe ejecutar el comando:

> survdiff(Surv(meses,censor2)~diabetes)

Obteniéndose el resultado:

Call:

survdiff(formula = Surv(meses, censor2) ~ diabetes)

	N	Observed	Expected	(O-E)∧2/E	(O-E)∧2/V
diabetes=0	212	49	56.46	0.986	8.6
diabetes=1	34	15	7.54	7.386	8.6

Chisq= 8.6 on 1 degrees of freedom, p= 0.00335

Y como p=0.00335 $<$ 0.05, se rechaza la hipótesis nula de igualdad de funciones de sobrevivencia (para un nivel de significación del 5%).

Capítulo 5

El modelo de regresión de Cox

El modelo de regresión de Cox (1972) es el modelo de regresión más utilizado para datos de supervivencia en el área médica.

El modelo de Cox posee la ventaja de que permite modelar covariables que dependen del tiempo, sin embargo, este tipo de modelaje no será abordado en este cursillo.

5.1 El modelo de Cox

En el modelo de regresión de Cox, el riesgo para el i-ésimo individuo se define mediante la siguiente expresión:

$$\lambda\left(t; Z_i\left(t\right)\right) = \lambda_0\left(t\right) e^{\beta' Z_i(t)}$$

donde $Z_i\left(t\right)$ es el vector de covariables para el i-ésimo individuo en el tiempo t.

El modelo de Cox establecido anteriormente se dice que es un modelo semiparamétrico debido a que incluye una parte paramétrica y

otra parte no paramétrica.

i) La parte paramétrica es $r_i(t) = e^{\beta' Z_i(t)}$, llamada puntaje de riesgo (risk score), y β es el vector de parámetros de la regresión.

ii) La parte no paramétrica es $\lambda_0(t)$ que es llamada función de riesgo base, es una función arbitraria y no especificada.

El modelo de regresión de Cox se llama también modelo de riesgos proporcionales debido a que el cociente entre el riesgo para dos sujetos con el mismo vector de covariables es constante en el tiempo, es decir:

$$\frac{\lambda(t;Z_i(t))}{\lambda(t;Z_j(t))} = \frac{\lambda_0(t)e^{\beta' Z_i(t)}}{\lambda_0(t)e^{\beta' Z_j(t)}} = \frac{e^{\beta' Z_i(t)}}{e^{\beta' Z_j(t)}}$$

Suponiendo que una muerte ha ocurrido en el tiempo t^* , entonces la verosimilitud de que la muerte le ocurra al individuo i-ésimo y no a otro individuo es:

$$L_i(\beta) = \frac{\lambda_0(t^*)r_i(t^*)}{\sum_j Y_j(t^*)\lambda_0(t^*)r_j(t^*)} = \frac{r_i(t^*)}{\sum_j Y_j(t^*)r_j(t^*)}$$

El producto de los términos de la expresión anterior $L(\beta) = \prod L_i(\beta)$ es llamada la verosimilitud parcial y fue introducida por Cox (1972).

La maximización de $\log(L(\beta))$ da una estimación para β sin necesidad de estimar el parámetro de ruido o función de riesgo base $\lambda_0(t)$

5.2 Contrastes de hipótesis para el modelo de Cox

Una vez que se ha ajustado un modelo de Cox, existen tres contrastes de hipótesis para verificar la significación del modelo, estos tests son asintóticamente equivalentes, pero no siempre sucede lo mismo en la práctica.

5.2.1 Test de razón de verosimilitud

El primero de los contrastes es el denominado test de razón de verosimilitud y es el que presenta una mayor confiabilidad. Este test se define como:

$$2\left\{\log\left(L\left(\beta_0\right)\right)-\log\left(L\left(\hat{\beta}\right)\right)\right\}$$

donde β_0 son los valores iniciales de los coeficientes y $\hat{\beta}$ es la solución luego de ajustar el modelo.

5.2.2 Test de Wald

El segundo de los contrastes es conocido como el test de Wald y es quizás el más natural debido a que proporciona un contraste por variables en vez de una medida de significación global. El estadístico de contraste se define mediante:

$$\left(\hat{\beta}-\beta_0\right)^{!}\hat{\Sigma}_{\hat{\beta}}^{-1}\left(\hat{\beta}-\beta_0\right)$$

donde $\hat{\Sigma}_{\beta}$ es la matriz de varianzas y covarianzas estimada.

5.2.3 Test de puntajes (score test)

El tercer contraste es el conocido como test de los puntajes, definido como$U'IU$, donde U es el vector de derivadas del $\log\left(L\left(\beta\right)\right)$ dado por:

$$U\left(\beta\right)=\sum_{i=1}^{n}\int_{0}^{\infty}\left[Z_i\left(t\right)-\bar{Z}\left(\beta,t\right)\right]dN_i\left(t\right)$$

I es la matriz de información dada por:

$$I\left(\beta\right)=\sum_{i=1}^{n}\int_{0}^{\infty}\frac{\sum_j Y_j(t)r_j(t)\left[Z_i(t)-\bar{Z}(\beta,t)\right]\left[Z_i(t)-\bar{Z}(\beta,t)\right]^{!}}{\sum_j Y_j(t)r_j(t)}dN_i\left(t\right)$$

y $\bar{Z}(\beta, t)$ es la media de las covariables para aquellos todavía en riesgo en el tiempo t, dada por:

$$\bar{Z}(\beta, t) = \frac{\sum_j Y_j(t) r_j(t) Z_i(t)}{\sum_i Y_i(t) r_i(t)}$$

5.3 Modelos de Cox estratificados

Una extensión del modelo de Cox permite obtener la estimación de los modelos para distintos grupos disjuntos o estratos. El modelo obtenido se conoce como modelo de Cox estratificado y está definido para el estrato j-ésimo como:

$$\lambda(t; Z_i(t)) = \lambda_j(t) e^{\beta' Z_i(t)}$$

Este modelo permite obtener la estimación del modelo en presencia de una variable de estratificación sobre la cual se desean obtener funciones de supervivencia por cada uno de los distintos grupos y probablemente poder estudiar la existencia o no de las funciones de sobrevivencia entre los grupos.

El modelo de Cox estratificado también constituye una de las maneras de corregir el modelo de Cox cuando no se cumple el supuesto de riesgos proporcionales para alguna de las covariables. En este caso suele correrse el modelo estratificando por la covariable que no cumple con el supuesto de riesgo proporcional. Este procedimiento permite corregir el sesgo en la estimación del parámetro que puede presentarse cuando se viola el supuesto de riesgo proporcional. Sin embargo, presenta una desventaja y es que no existe ningún β que permita estimar el efecto de la covariable de estratificación.

5.4 Estudio de residuos en el análisis de supervivencia

Una de las ventajas que han surgido del enfoque del análisis de supervivencia es la posibilidad de efectuar análisis de residuos (Andersen et al., 1993, Fleming y Harrington, 1991, Therneau y Grambsch, 2000, Therneau et al., 1990).

Los residuos se pueden utilizar para:
1. Descubrir la forma funcional correcta de un predictor continuo.
2. Identificar los sujetos que están pobremente predichos por el modelo.
3. Identificar los puntos o individuos de influencia.
4. Verificar el supuesto de riesgo proporcional.

Existen cuatro tipos de residuos de interés en el modelo de Cox: los residuos de martingala, los de desvíos (deviances), los de puntaje (score) y los de Schoenfeld. De estos cuatro residuos pueden derivarse otros dos: los dfbetas y los residuos escalados de Schoenfeld. A continuación explicaremos brevemente cada uno de estos residuos.

5.4.1 Residuos de martingala

Los residuos de martingala se definen como:

$$\hat{M}_i(t) = N_i(t) - \hat{E}_i(t) = N_i(t) - \int_0^t Y_i(s)\, e^{\beta' Z_i(s)} d\hat{\Lambda}_0(\beta, s)$$

donde $\hat{\Lambda}_0(\beta, s)$ es el estimador del riesgo base de Breslow (o de Tsiatis o de Nelson y Aalen) definido como:

$$\hat{\Lambda}_0(\beta, s) = \int_0^t \frac{\sum_{i=1}^{n} dN_i(s)}{\sum_{i=1}^{n} Y_i(s) e^{\beta' Z_i(s)}}$$

y están basados en la martingala de un proceso de conteo para el i-ésimo individuo, $M_i(t) = N_i(t) - E_i(t)$, definida mediante:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\, e^{\beta' Z_i(s)} \lambda_0(s) ds$$

Los residuos de martingala son muy asimétricos y con una cola muy larga hacia la derecha, particularmente para datos de supervivencia para un solo evento.

Los residuos de martingala se usan para estudiar la forma funcional de una covariable.

5.4.2 Residuos de desvíos (deviances)

Los residuos de desvíos se obtienen mediante una transformación de normalización de los desvíos de martingala y son similares en forma a los residuos de desvíos (deviances) en la regresión de Poisson.

Los residuos de desvíos se definen de la manera siguiente: si todas las covariables son fijas en el tiempo, los residuos toman la forma:

$$d_i = \text{signo}\left(\hat{M}_i\right) * \sqrt{-\hat{M}_i - N_i \log\left(\left(N_i - \hat{M}_i\right)/N_i\right)}$$

Una expansión de Taylor de un término muestra que:

$$d_i \approx \frac{N_i - \hat{E}_i}{\sqrt{\hat{E}_i}}$$

que es formalmente equivalente a los residuos de Pearson de los modelos lineales generalizados.

Los residuos de desvíos se utilizan para la detección de valores atípicos (outliers).

5.4.3 Residuos de puntajes (scores)

Los residuos de puntajes se definen como:

$$U_{ij} = U_{ij}\left(\hat{\beta}, \infty\right)$$

donde $U_{ij}\left(\beta, t\right), j = 1, \cdots, p$ son las componentes del vector fila de longitud p obtenido a través del proceso de puntaje para el i-ésimo individuo:

$$U_i\left(\beta\right) = \int_0^t \left[Z_i\left(t\right) - \bar{Z}\left(\beta, t\right)\right] dN_i\left(t\right)$$

Los residuos de puntajes se utilizan para verificar la influencia individual y para la estimación robusta de la varianza.

5.4.4 Residuos de Schoenfeld

Los residuos de Schoenfeld (1982) se definen como la matriz:

$$s_{ij}\left(\beta\right) = Z_{ij}\left(t_i\right) - \bar{Z}_j\left(\beta, t_i\right)$$

con una fila por muerte y una columna por covariable, donde i y t_i son los individuos y el tiempo de ocurrencia del evento respectivamente.

Los residuos de Schoenfeld son útiles para la verificación del supuesto de riesgo proporcional en el modelo de Cox.

5.5 Interpretación del modelo de Cox

La interpretación del modelo de Cox no se hace directamente a través de su coeficiente estimado sino del exponencial de la estimación del coeficiente estimado, $\exp\left(\hat{\beta}\right)$.

Para variables dicotómicas $\exp\left(\hat{\beta}\right)$ es un estimador de la razón

de riesgos (hazard ratio) y se interpreta como la cantidad de riesgo que se tiene con la presencia de cada covariable en relación a la ausencia del resto de las la covariables.

Los intervalos de confianza del 95% para $\exp\left(\hat{\beta}\right)$ se obtienen mediante:

$$\exp\left(\hat{\beta} \pm 1.96 ee\left(\hat{\beta}\right)\right)$$

donde $ee\left(\hat{\beta}\right)$ es el error estándar de $\hat{\beta}$.

Para el caso de covariables continuas, $\exp\left(\hat{\beta}\right)$ representa la razón de riesgos (hazard ratio) al incrementar en una unidad la covariable.

Resulta más interesante estimar la razón de riesgos al incrementar la covariable en c unidades y esto se hace mediante $\exp\left(c\hat{\beta}\right)$, siendo su intervalo de confianza del 95% de la forma:

$$\exp\left(c\hat{\beta} \pm 1.96\,|c|\, ee\left(\hat{\beta}\right)\right)$$

Para una explicación más detallada puede verse Hosmer y Lemeshow (1999).

5.6 Ejemplo 5.1

Este ejemplo es continuación del del ejemplo 4.1. En este capí se presentará lo relativo al modelo de Cox.

5.6.1 Ajuste del modelo de Cox

Ajustemos un modelo de Cox con diabetes, edad e índice de quetelet como covariables y asignemos el ajuste al objeto con nombre cox1, esto lo hacemos, mediante la instrucción:

> cox1< −coxph(Surv(meses,censor2)∼diabetes + edad + quetelet, data=dpa, na.action=na.exclude)

Y al ejecutar el comando print(cox1) (o simplemente cox1), con lo que obtenemos el siguiente resultado:

Call:
coxph(formula = Surv(meses, censor2)∼diabetes + edad + quetelet, data = dpa, na.action = na.exclude)

	coef	exp(coef)	se(coef)	z	p
diabetes	0.5491	1.732	0.3208	1.71	0.0870
edad	0.0315	1.032	0.0097	3.25	0.0011
quetelet	-0.0969	0.908	0.0389	-2.49	0.0130

Likelihood ratio test=18.8 on 3 df, p=0.000308 n=233 (13 observations deleted due to missing)

Que es una salida en donde la significación del modelos puede verificarse sólo a través del método de la razón de verosimilitud. Una salida más completa se presenta mediante la ejecución del comando summary(cox1), la cual tiene la forma:

Call:
coxph(formula = Surv(meses, censor2) ∼ diabetes + edad + quetelet, data = dpa, na.action = na.exclude)

n=233 (13 observations deleted due to missing)

	coef	exp(coef)	se(coef)	z	p
diabetes	0.5491	1.732	0.3208	1.71	0.0870
edad	0.0315	1.032	0.0097	3.25	0.0011
quetelet	-0.0969	0.908	0.0389	-2.49	0.0130

	exp(coef)	exp(-coef)	lower .95	upper .95
diabetes	1.732	0.577	0.923	3.25
edad	1.032	0.969	1.013	1.05
quetelet	0.908	1.102	0.841	0.98

Rsquare= 0.077 (max possible= 0.892)

Likelihood ratio test= 18.8 on 3 df, p=0.000308
Wald test = 19.4 on 3 df, p=0.000229
Score (logrank) test = 19.8 on 3 df, p=0.000184

Con la cual podemos concluir que el modelo es significativo cualquiera de los tres criterios (test de razón de verosimilitud, test de Wald y test de los puntajes (score o logrank)).

La salida anterior también nos permite verificar la significación de cada uno de los coeficientes correspondientes a las covariables, observándose que el coeficiente correspondiente a diabetes es significativo al 10%, el de la edad lo es al 1% y el del índice de Quetelet lo es al 5%.

Otra información importante, obtenida directamente a través de salida anterior es la estimación de los riesgos relativos (a partir de los exp(coef)), con los cuales podemos decir que la presencia de diabetes hace que la muerte tenga un riesgo de 1.732 veces el riesgo de muerte de los no diabéticos. En cuanto a la edad, una persona con una edad determinada tiene 1.032 veces el riesgo de morir en relación a una persona una año menor, para esta covariable, la interpretación pudiera también darse en relación a quinquenios obteniéndose un riesgo relativo de 1.17=exp(5*0.0315). Finalmente, al aumentar el índice de Quetelet en una unidad el riesgo se hace 0.908 veces que la del menor valor.

Mediante el comando:

> summary(survfit(cox1))

Puede obtenerse la función de sobrevivencia ajustada mediante el modelo de Cox, cuya salida es :

Call: survfit.coxph(object = cox1)

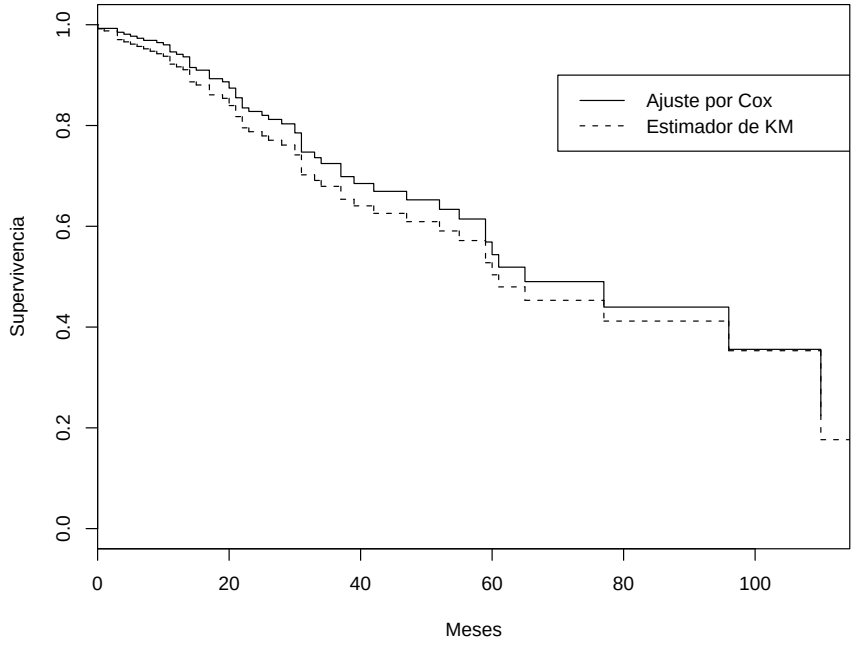
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	233	2	0.993	0.00516	0.983	1.000
3	220	2	0.985	0.00751	0.970	1.000
4	215	1	0.981	0.00847	0.965	0.998
5	209	1	0.977	0.00938	0.959	0.996
6	203	1	0.973	0.01024	0.953	0.993
7	196	1	0.969	0.01109	0.947	0.991
9	188	1	0.965	0.01195	0.941	0.988
10	183	1	0.960	0.01278	0.935	0.985
11	175	3	0.946	0.01520	0.917	0.976
12	166	1	0.941	0.01597	0.910	0.973
13	156	1	0.936	0.01678	0.904	0.970
14	148	4	0.915	0.01996	0.877	0.955
15	142	1	0.910	0.02069	0.870	0.951
17	133	3	0.893	0.02297	0.849	0.939
19	122	1	0.887	0.02378	0.841	0.935
20	117	2	0.874	0.02539	0.826	0.925
21	113	3	0.855	0.02768	0.802	0.911
22	108	3	0.835	0.02992	0.778	0.896
23	102	1	0.828	0.03068	0.770	0.890
25	92	1	0.820	0.03156	0.760	0.884
26	88	1	0.812	0.03249	0.751	0.878
28	79	1	0.803	0.03356	0.740	0.872
30	76	2	0.785	0.03566	0.718	0.858
31	73	4	0.747	0.03959	0.673	0.829
33	62	1	0.736	0.04074	0.660	0.820
34	58	1	0.725	0.04195	0.647	0.812
37	52	2	0.699	0.04474	0.616	0.792
39	49	1	0.685	0.04609	0.600	0.781
42	42	1	0.669	0.04781	0.582	0.770
47	37	1	0.652	0.04975	0.562	0.758
52	32	1	0.634	0.05207	0.539	0.744
55	30	1	0.614	0.05423	0.517	0.730
59	25	2	0.569	0.05965	0.463	0.699
60	21	1	0.544	0.06245	0.434	0.681
61	20	1	0.519	0.06474	0.406	0.663
65	17	1	0.490	0.06759	0.374	0.642
77	10	1	0.440	0.07749	0.311	0.621
96	6	1	0.356	0.09776	0.208	0.610
110	3	1	0.224	0.12059	0.078	0.643

También puede obtenerse la gráfica correspondiente, pero en esta ocasión graficaremos la función de sobrevivencia obtenida mediante el estimado de Kaplan y Meier y la obtenida mediante el modelo de Cox, esto lo podemos hacer mediante los comandos:

```
> plot(survfit(cox1),conf.int=FALSE,main="Gráfico
No. 3. Comparación del ajuste del modelo de Cox \n y el estimador
de KM",xlab="Meses",ylab="Supervivencia")
> lines(km1,lty=2)
> legend(70,0.9,legend=c("Ajuste por Cox","Estimador de KM"),
lty=c(1,2))
```

Obteniéndose la gráfica:

Gráfico No. 3. Comparación del ajuste del modelo de Cox y el estimador de KM



Pudiéndose observar que el ajuste del modelo de Cox es sistemáticamente superior a la función de sobrevivencia de Kaplan y Meier.

5.6.2 Verificación de los supuestos del modelo de Cox

Supuesto de riesgos proporcionales:

El supuesto de riesgos proporcionales puede ser verificado mediante el contraste de hipótesis generado mediante el comando:

```
> cox.zph(cox1)
```

del cual se obtiene la salida:

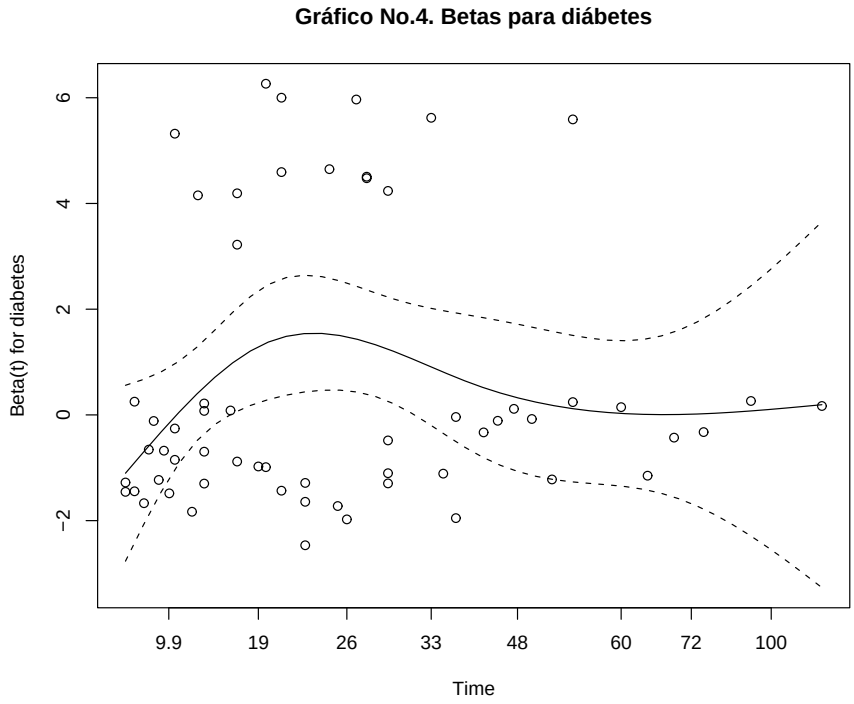
	rho	chisq	p
diabetes	0.0357	0.0808	0.776
edad	0.1165	1.0519	0.305
quetelet	-0.0540	0.2278	0.633
GLOBAL	NA	1.3791	0.710

De donde se concluye de que no existe evidencia significativa al 5% de que se viole el supuesto de riesgos proporcionales, ni desde el punto de vista global, ni para cada covariable.

Para cada una de las covariables también pueden obtenerse los gráficos para los betas. Para la Variable diabetes, el gráfico correspondiente se obtiene mediante el comando:

```
> plot(cox.zph(cox1),var=1,main="Gráfico No.4. Betas para diábetes")
```

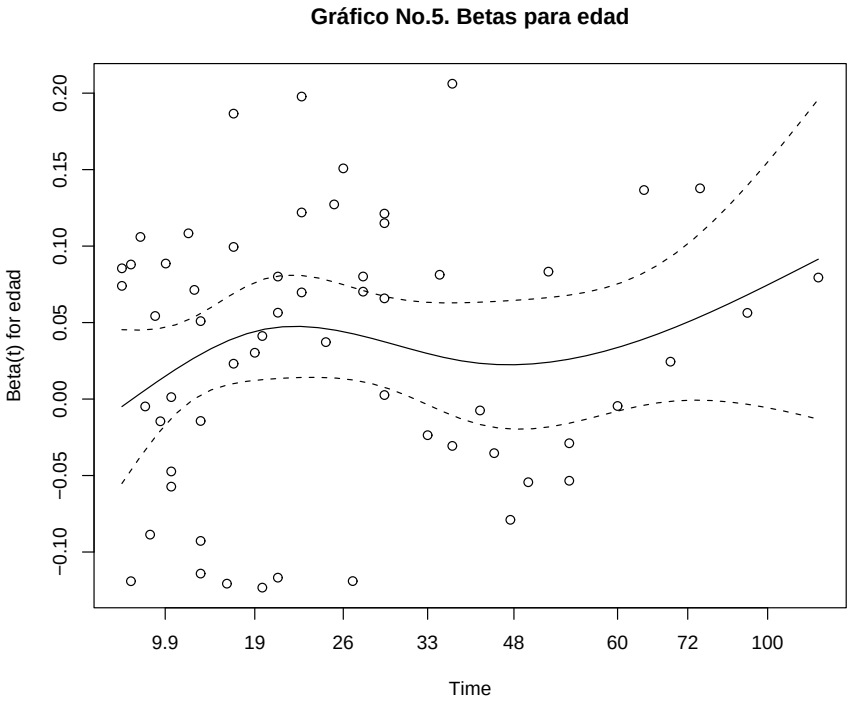
Y toma la forma:



Para la variable edad, el código se escribe como:

> plot(cox.zph(cox1),var=2,main="Gráfico No.5. Betas para edad")

Y el gráfico es:

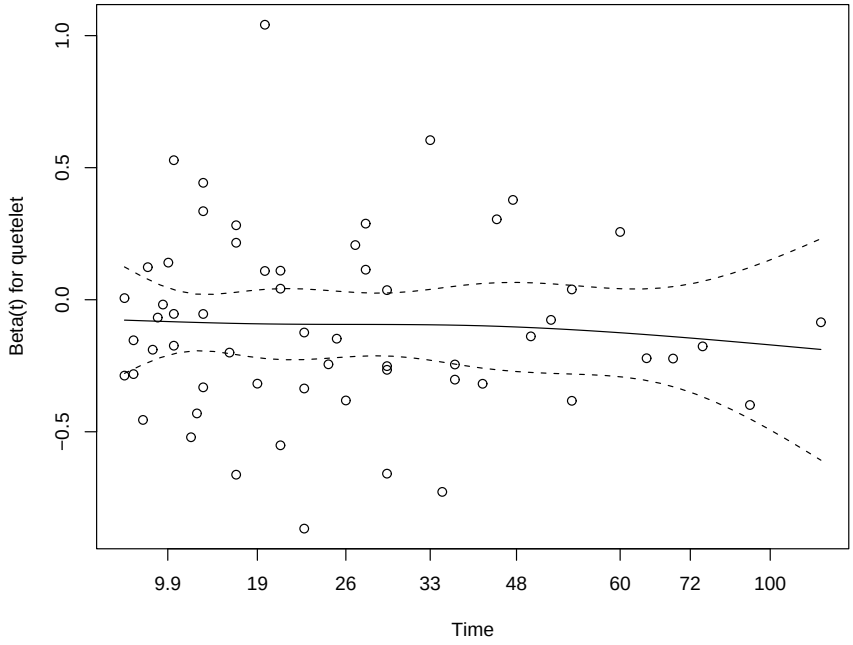


Y para la variable índice de Quetelet, el comando es:

> plot(cox.zph(cox1),var=3,main="Gráfico No.6. Betas para índice de Quetelet")

Y el gráfico es:

Gráfico No.6. Betas para índice de Quetelet



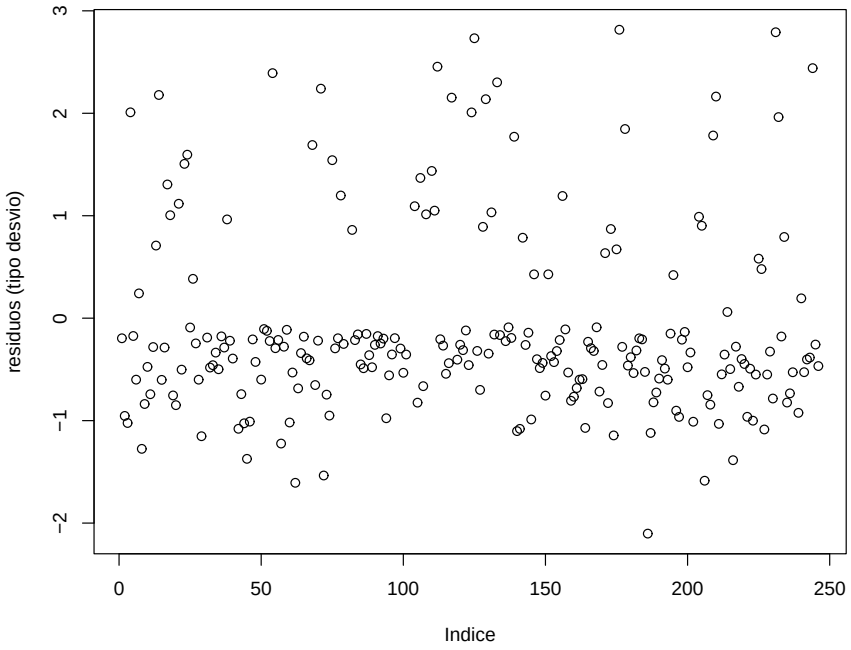
Residuos tipo deviance:

Los residuos tipo deviance pueden generarse a través del comando:

> plot(resid(cox1,type="deviance"),xlab="indice",ylab="residuos (tipo desvio)", main="Gráfico No. 7. Residuos (tipo deviance)")

Los cuales generan el gráfico:

Gráfico No. 7. Residuos (tipo deviance)



Y en cual se evidencia que no existe ningún individuo que esté influenciando en el ajuste del modelo.

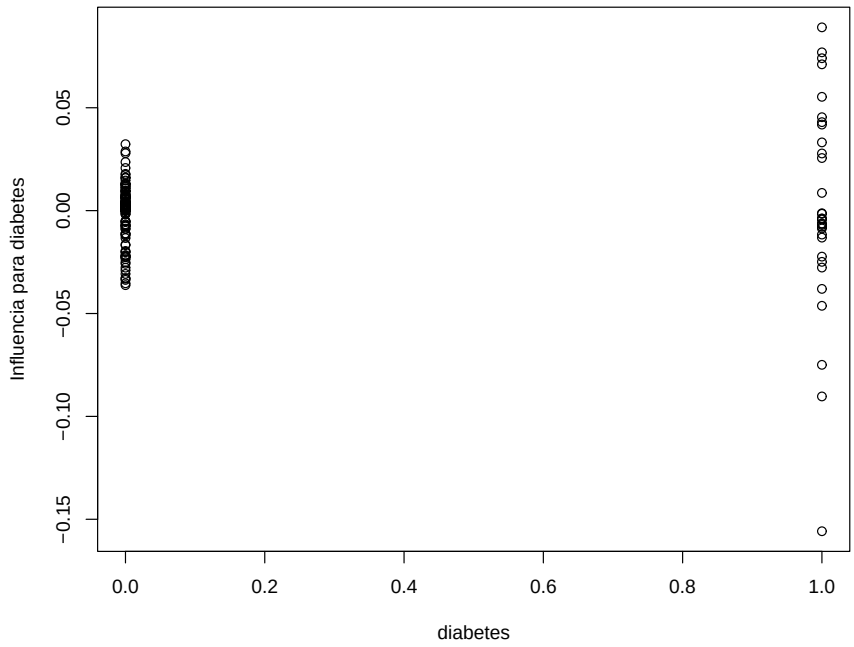
Gráficos de influencias sobre la estimación de cada coeficiente:

Estos gráficos se obtienes utilizando los residuos dfbetas. Para el caso de diabetes, el gráfico se genera a través de los comandos:

```
> rr< −resid(cox1,type="dfbeta")
> attach(dpa) > plot(diabetes, rr[,1], xlab="diabetes", ylab="Influencia
para diabetes", main="Gráfico No. 8. Gráfico de influencias
para diabetes")
```

Obteniéndose el gráfico:

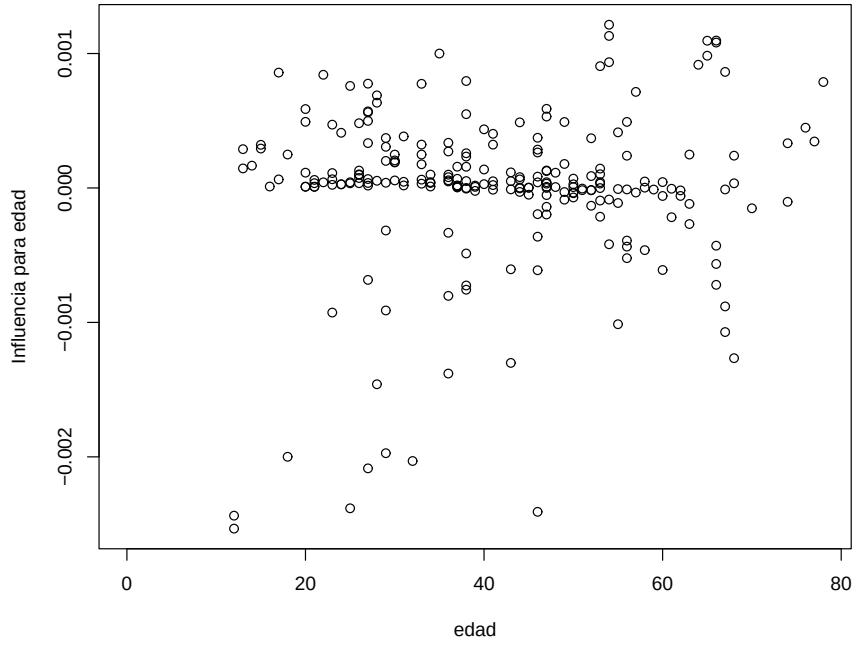
Gráfico No. 8. Gráfico de influencias para diabetes



Para la variable edad, el gráfico se obtiene mediante el comando:
> plot(edad,rr[,2],xlab="edad",ylab="Influencia para edad", main= "Gráfico No. 9. Gráfico de influencias para edad")

del cual se obtiene el siguiente gráfico:

Gráfico No. 9. Gráfico de influencias para edad

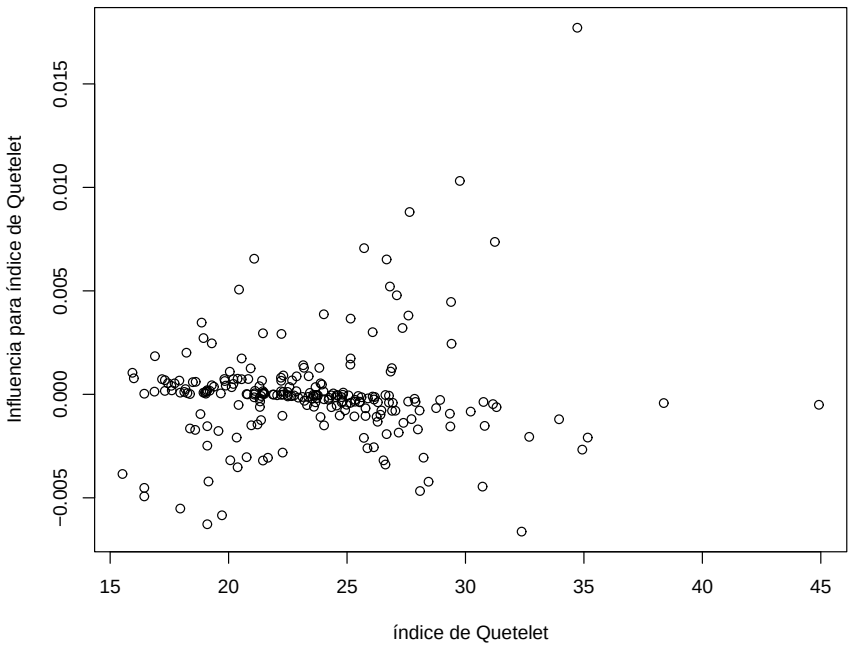


Finalmente, para el índice de Quetelet, se obtiene mediante el comando:

```
> plot(quetelet,rr[,3],xlab="índice de Quetelet",ylab="Influencia
para índice de Quetelet",main="Gráfico No. 10. Gráfico de
influencias \n el índice de Quetelet")
```

Con el cual se genera el gráfico:

Gráfico No. 10. Gráfico de influencias el índice de Quetelet



Pudiéndose observar que existe un individuo que esta influenciando sobre la estimación del coeficiente correspondiente al índice de Quetelet.

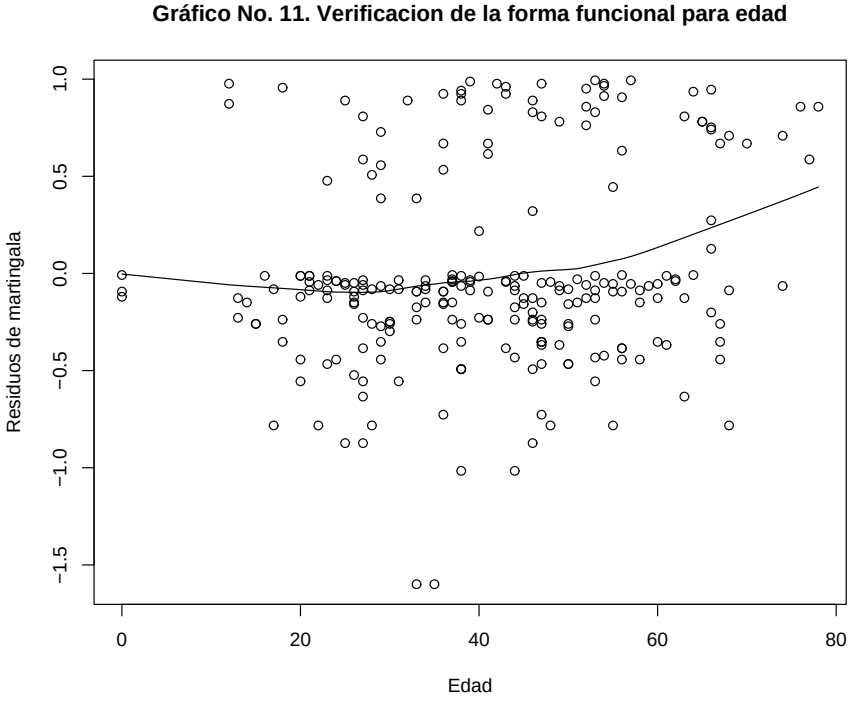
Forma funcional de las variables continuas

La adecuacidad de la forma funcional para la variable edad puede ser verificada a través de los residuos de martingala, mediante el siguiente código:

```
> cox1.0< −coxph(Surv(meses,censor2)∼1,data=dpa,
na.action=na.exclude)
> rr< −resid(cox1.0)
> plot(dpa$edad,rr,xlab="Edad",ylab="Residuos de martingala",
```

main="Gráfico No. 11. Verificacion de la forma funcional para edad")
> lines(lowess(dpa$edad,rr,iter=0))

Obteniéndose el gráfico:



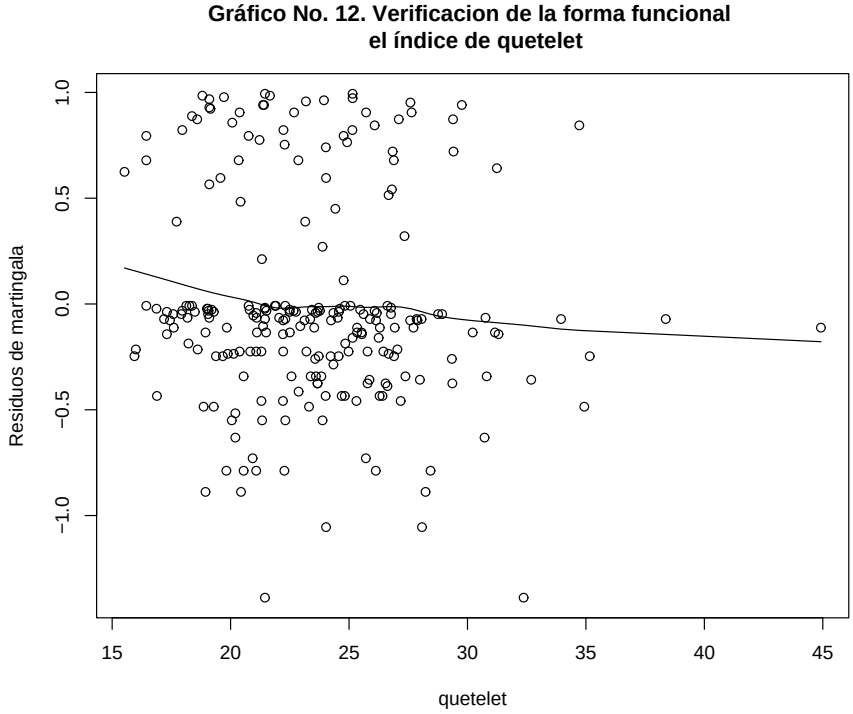
Observándose una forma funcional adecuada para la variable edad.

Para el caso de la variable índice de Quetelet, la adecuacidad de la forma funcional se puede verificar mediante los siguientes comandos:

> dpa.nq< − na.omit(dpa[, c("meses","censor2","diabetes", "edad", "quetelet")])

```
> cox.nq.0< −coxph(Surv(meses,censor2) 1,data=dpa.nq)
> rr< −resid(cox.nq.0)
> plot(dpa.nq$quetelet,rr,xlab="quetelet",ylab="Residuos de
martingala", main="Gráfico No. 12. Verificacion de la forma
funcional \n el índice de quetelet")
> lines(lowess(dpa.nq$quetelet,rr,iter=0)
```

el cual genera el siguiente gráfico:



En el que se observa que la forma funcional para el índice de Quetelet parece ser adecuada.

Capítulo 6

Modelos de regresión paramétricos

Otra forma de modelar las funciones de riesgo es a través de los modelos de regresión paramétricos. Los modelos de regresión paramétricos se basan en diversas familias de distribuciones comúnmente utilizadas, principlmente en Ańalisis de Confiabilidad y son de uso com'n en el área de la industria.

Existe un conjunto de distribuciones que estaán basadas en parametros de localización y escala. Algunas distribuciones de este tipo han ampliamente estudiadas en la literatura estadística como lo son: la distribucción exponencial, la Weibull, la Normal o Gaussiana, la Logística, la Lognormal y la Loglogística. Los modelos de regresin basados en estas distribuciones se encuetran disponibles en casi todos los paquetes estadísticos, incluyendo el Lenguaje R.

Existen otras familas de distribuciones basadas en paramétros de localizacin y escala, cuya regresión no está disponible facilmente en los software estadísticos pero que con algunas manipulaciones pueden ser modeladas a través de algunas herramientas, entre las

que se encuentra el Lenguaje R. Estas distribuciones incluyen: la distribución valor extremo más pequeño y la distribución valor extremo más grande.

Otras familas de distribuciones paramétricas utilizadas en modelos de regresión paramétricos son: la distribución Gamma, la Gamma generalizada, la Gamma generalizada extendida, le F gereralizada, la Gaussiana inversa, la de Birnbaum-Saunders, la Gompertz-Makeham, entre otras.

Para un estudio detallado puede consultarse los textos de Miller (1982), Kalbfleisch y Prentice (2002) o Lawless (2003). Un estudio más detallado es presentado en Meeker y Escobar (1998).

Las distribuciones paramétricas utilizadas en los modelos de regresión estudiadas en el análisis de sobrevivencia en en confiabilidad, también son abordadas en detalle en textos relacionados con distribuciones, veáse por ejemplo los textos de Johnson, Kotz y Balakrishnan (1994, 1995).

6.1 Características de algunos modelos paramétricos

En esta sección describiremos características de algunas distribuciones de uso común en modelos paramétricos, dedicándonos al estudio de las familias de distribuciones de localización y escala. Los gráficos asociados a estas distribuciones pueden verse en el texto de Meeker y Escobar, igualmente pueden verse las características de otras distribuciones.

6.1.1 Distribuciones de localización y escala.

Una variable aleatoria Y pertenece a la familia de distribuciones de localización y escala si su función de distribución puede ser expresada como:

$$F\left(y;\mu,\sigma\right)=P\left(Y\leq y\right)=\Phi\left(\tfrac{y-\mu}{\sigma}\right)$$

donde:

Φ no depende de un parámetro desconocido.

$-\infty<\mu<\infty$ es un parámetro de localización.

σ es un parámetro de escala.

La familia de distribuciones de localización y escala son importantes entre otras cosas debido a que la inferencia y el software computacional de estas distribuciones pueden ser aplicados con relativa facilidad a cualquier miembro de la familia.

Algunos miembros de esta familia se describen en las siguientes subsecciones.

Distribución exponencial de dos parámetros:

Se dice que la variable aleatoria T se distribuye como una exponencial de parámetro de localización γ y parámetro de escala $\frac{1}{\lambda}>0$, coincidiendo este con la esperanza, si sus funciones de distribución, densidad y riesgo toman las expresiones:

$$F\left(t;\lambda,\gamma\right)=1-e^{-(t-\gamma)\lambda}$$

$$f\left(t;\lambda,\gamma\right)=\lambda e^{-(t-\gamma)\lambda}\text{ y,}$$

$$h\left(t;\lambda,\gamma\right)=\lambda$$

Es fácil observar que si $\gamma = 0$, la distribución anterior es la exponencial de un parámetro, descrita en la siguiente subsección.

Distribución exponencial de un parámetro:

Se dice que la variable aleatoria T se distribuye como una exponencial de parámetro $\lambda > 0$ si su función de densidad toma la expresión:

$$f(t;\lambda) = \lambda e^{-\lambda t}$$

de donde, la función de sobrevivencia toma la forma:

$$S(t;\lambda) = \int_t^\infty f(u)\,du = e^{-\lambda t}$$

y la función de riesgo es:

$$\lambda(t;\lambda) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

de donde se puede asumir que el modelo exponencial es de riesgo constante.

por lo tanto, la función de riesgo acumulada es:

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

Distribución Weibull:

El modelo Weibull es una generalización del modelo exponencial. Se dice que la variable aleatoria T se distribuye como una exponencial de parámetros $\alpha > 0$ y $\lambda > 0$ si su función de densidad toma la expresión:

$$f(t;\lambda,\alpha) = \alpha\lambda(\lambda t)^{\alpha-1} e^{-(\lambda t)^\alpha}$$

por lo tanto, la función de sobrevivencia es:

$$S\left(t;\lambda,\alpha\right)=\int_{t}^{\infty}f\left(u\right)du=e^{-\left(\lambda t\right)^{\alpha}}$$

y la función de riesgo es:

$$\lambda\left(t;\lambda,\alpha\right)=\frac{f(t)}{S(t)}=\frac{\alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}}}{e^{-(\lambda t)^{\alpha}}}=\alpha\lambda\left(\lambda t\right)^{\alpha-1}$$

y la función de riesgo acumulada toma la forma:

$$\Lambda\left(t;\lambda,\alpha\right)=\int_{0}^{t}\lambda\left(u\right)du=\left(\lambda t\right)^{\alpha}$$

Puede notarse que la distribución de Weibull pertenece a la familia de localización y escala con parámetro de escala $\frac{1}{\lambda}$ y α es un parámetro de forma.

Distribución normal:

Se dice que la variable aleatoria T se distribuye como una normal de parámetros μ y σ^2. Las funciones de distribución y densidad toman la forma:

$$F\left(t;\mu,\sigma^{2}\right)=\Phi_{nor}\left(\tfrac{t-\mu}{\sigma}\right)$$

$$f\left(t;\mu,\sigma^{2}\right)=\tfrac{1}{\sigma}\phi_{nor}\left(\tfrac{t-\mu}{\sigma}\right)$$

donde:

$\phi_{nor}\left(z\right)=\left(1\Big/\sqrt{2\pi}\right)e^{-\left(z^{2}/2\right)}$ es la función de densidad de la normal estándar y,

$\Phi_{nor}=\int_{-\infty}^{z}\phi_{nor}\left(w\right)dw$ es la función de distribución de la normal estándar.

Puede observarse que para la distribución normal, $-\infty<\mu<\infty$ es un parámetro de localización y $\sigma>0$ es un parámetro de escala.

Distribución Lognormal:

Decimos que la variable aleatoria T se distribuye como una Lognormal de parámetros μ y σ^2 si su logaritmo se distribuye como una normal de parámetros μ y σ^2.

Las funciones de distribución y de densidad de la distribución Lognormal toman las expresiones:

$$F\left(t;\mu,\sigma^2\right) = \Phi_{nor}\left[\frac{\log(t)-\mu}{\sigma}\right] \text{ y,}$$

$$f\left(t;\mu,\sigma^2\right) = \frac{1}{\sigma t}\phi_{nor}\left[\frac{\log(t)-\mu}{\sigma}\right]$$

En esta distribución, la mediana (e^{μ}) es una parámetro de escala y σ es una parámetro de forma.

Distribución valor extremo más pequeño:

Se dice que una variable aleatoria T se distribuye como una distribución de valor extremo más pequeño de parámetros μ y σ si sus funciones de distribución, de densidad y de riesgo toman la forma:

$$F\left(t;\mu,\sigma\right) = \Phi_{sev}\left(\frac{t-\mu}{\sigma}\right),$$

$$f\left(t;\mu,\sigma\right) = \frac{1}{\sigma}\phi_{sev}\left(\frac{t-\mu}{\sigma}\right) \text{ y,}$$

$$h\left(t;\mu,\sigma\right) = \frac{1}{\sigma}e^{\left(\frac{t-\mu}{\sigma}\right)}$$

donde:

$\Phi_{sev}\left(z\right) = 1 - e^{-e^z}$ es la función de distribución de la distribución valor extremo más pequeño estándar.

$\phi_{sev}\left(z\right) = e^{(z-e^z)}$ es la función de densidad de la distribución valor

extremo más pequeño estándar.

Puede observarse que para la distribución valor extremo más pequeño, $-\infty < \mu < \infty$ es un parámetro de localización y $\sigma > 0$ es un parámetro de escala.

Distribución valor extremo más grande:

Se dice que una variable aleatoria T se distribuye como una distribución de valor extremo más grande de parámetros μ y σ si sus funciones de distribución, de densidad y de riesgo toman la forma:

$$F\left(t;\mu,\sigma\right)=\Phi_{lev}\left(\tfrac{t-\mu}{\sigma}\right),$$

$f\left(t;\mu,\sigma\right)=\frac{1}{\sigma}\phi_{lev}\left(\frac{t-\mu}{\sigma}\right)$ y,

$$h\left(t;\mu,\sigma\right)=\frac{e^{-\left(\frac{t-\mu}{\sigma}\right)}}{\sigma e^{\left[e^{-\left(\frac{t-\mu}{\sigma}\right)}-1\right]}}$$

donde:

$\Phi_{lev}\left(z\right)=e^{-e^{z}}$ es la función de distribución de la distribución valor extremo más grande estándar.

$\phi_{lev}\left(z\right)=e^{\left(-z-e^{z}\right)}$ es la función de densidad de la distribución valor extremo más grande estándar.

Puede observarse que para la distribución valor extremo más grande, $-\infty < \mu < \infty$ es un parámetro de localización. Y $\sigma > 0$ es un parámetro de escala.

Distribución Logística:

Se dice que una variable aleatoria T se distribuye como una distribución Logística de parámetros μ y σ si sus funciones de dis-

tribución, de densidad y de riesgo toman la forma:

$$F\left(t;\mu,\sigma\right)=\Phi_{\log is}\left(\tfrac{t-\mu}{\sigma}\right),$$

$$f\left(t;\mu,\sigma\right)=\tfrac{1}{\sigma}\phi_{\log is}\left(\tfrac{t-\mu}{\sigma}\right) \text{ y,}$$

$$h\left(t;\mu,\sigma\right)=\tfrac{1}{\sigma}\Phi_{\log is}\left(\tfrac{t-\mu}{\sigma}\right)$$

donde:

$\Phi_{\log is}\left(z\right)=\frac{e^{z}}{1+e^{z}}$ es la función de distribución de la distribución logística estándar.

$\phi_{\log is}\left(z\right)=\frac{e^{z}}{(1+e^{z})^{2}}$ es la función de densidad de la distribución logística estándar.

Puede observarse que para la distribución logística, $-\infty<\mu<\infty$ es un parámetro de localización. Y $\sigma>0$ es un parámetro de escala.

Distribución Loglogística:

Se dice que la variable aleatoria T se distribuye como una distribución Loglogística de parámetros μ y σ si $\log\left(T\right)$ se distribuye como una logística de parámetros μ y σ y, sus funciones de distribución, de densidad y de riesgo toman la forma:

$$F\left(t;\mu,\sigma\right)=\Phi_{\log is}\left(\tfrac{\log(t)-\mu}{\sigma}\right),$$

$$f\left(t;\mu,\sigma\right)=\tfrac{1}{\sigma t}\phi_{\log is}\left(\tfrac{\log(t)-\mu}{\sigma}\right) \text{ y,}$$

$$h\left(t;\mu,\sigma\right)=\tfrac{1}{\sigma t}\Phi_{\log is}\left(\tfrac{\log(t)-\mu}{\sigma}\right)$$

En la distribución Loglogística, la mediana (e^{μ}) es una parámetro de escala y σ es una parámetro de forma.

6.2 Estimación de los modelos paramétricos

Para el caso de que no exista censura, el problema de estimación de los parámetros de los modelos paramétricos se hace a través de los métodos clásicos, entre los que destaca el método de máxima verosimilitud.

Para el caso de presencia de censura, hay una parte de la verosimilitud que es aportada por los eventos observados y otra por los datos censurados y se utiliza también el método de máxima verosimilitud los cuales pueden ser resueltos a través del método iterativos, entre los que se encuentran, el método de Newton y Raphson y el método de la cantidad de información de Fisher (Miller, 1982, Lawless, 2003).

6.2.1 Caso general

Para obtener la estimación máximo verosímil se asume que para cada observación i, el par (y_i, δ_i) tiene la verosimilitud:

$$L(y_i, \delta_i) = \begin{cases} f(y_i) & \text{si } \delta_i = 1 \text{ (evento observado)} \\ S(y_i) & \text{si } \delta_i = 0 \text{ (dato censurado)} \end{cases}$$

y la verosimilitud de la nuestra aleatoria completa de tamaño n viene dada por :

$$L = L(y_i, \cdots, y_n, \delta_i, \cdots, \delta_n) = \prod_{i=1}^{n} L(y_i, \delta_i) = \left(\prod_{u} f(y_i)\right)\left(\prod_{c} S(y_i)\right)$$

donde los subíndices u y c representan los eventos observados y los datos censurados de la muestra aleatoria.

6.2.2 Estimación para las distribuciones de localización y escala

Para el caso de la familia de distribuciones de localización y escala, la maximización se hace a través de la expresión de la función de

verosimilitud de acuerdo a la siguiente expresión:

$$L(\mu,\sigma) = \prod_{i=1}^{n} \left[\frac{1}{\sigma}\phi\left(\frac{t_i-\mu}{\sigma}\right)\right]^{\delta_i} \left[1-\Phi\left(\frac{t_i-\mu}{\sigma}\right)\right]^{1-\delta_i}$$

Para más detalles puede verse el texto de Meeker y Escobar (1998).

6.3 Identificación del modelo paramétrico más adecuado

La identificacón del modelo paramétrico más adecuado es una asunto que requiere de cierto entrenamiento, en muchos caso se haría comparando la función de riesgo empírica con la función de riesgo teórica (pueden usarse también las funciones de distribuciones o las funciones de densidad). Estas fuciones se encuentran graficadas en algunos libros, como por ejemplo el de Meeker y Escobar (1998).

Sin embargo, estas comparaciones visuales son a menudo riesgosas y en algunos casos no puede identificarse una única distribución requiriéndose pruebas más refinadas como las de bondad de ajuste o el ploteo de probabilidad como el presentado en Meeker y Escobar (1998). Lamentablemente este tipo de gráficos no ha sido implementado en el Lenguaje R.

En S-PLUS (Insightful Corporation, 2001) existe una libreria llamada SLIDA desarrollada por Meeker y Escobar (1998), que permite construir con facilidad este tipo de gráficos.

6.3.1 Algunos gráficos que permiten identificar modelos paramétricos

Existen algunos gráficos sencillos que permiten identificar algunos modelos paramétricos. A continuación se presentan los más conocidos:

Modelo exponencial:

El modelo exponencial puede identificarse si al graficar la función de riesgo estimada $\hat{\lambda}(t)$ versus el tiempo t se observa aproximadamente una líinea recta horizontal.

Modelo Weibull

El modelo Weibull puede identificarse al:

i) Observar una línea recta que corta en el origen al graficar $-\log \hat{S}(t)$ versus el tiempo t, donde $\hat{S}(t)$ es la función de sobrevivencia estimada.

ii) Obtener una línea recta al graficar $\log\left[-\log \hat{S}(t)\right]$ versus el tiempo $\log(t)$.

Modelo Lognormal

El modelo Lognormal puede identificarse al:

i) Obtener una línea recta al graficar $\Phi^{-1}\left[1-\hat{S}(t)\right]$ versus $\log(t)$, donde $\Phi()$ es la funcíon de distribución de una normal estándar.

ii) Obtener una línea recta al graficar $\log\left[\left(1-\hat{S}(t)\big/\hat{S}(t)\right)\right]$ versus $\log(t)$.

Modelo Loglogístico

El modelo Loglogístico puede observarse al obtener una línea recta al graficar $Logit\left[\hat{S}(t)\right]$ versus $\log(t)$.

Para más detalles puede verse los textos de Allison (1995) O Miller (1982).

Criterio de los residuos de Cox y Snell

Collett (2003) propone la técnica de análisis de residuos de Cox y Snell, definidos como:

$$e_i = -\log S\left(t_i|\mathbf{x}_i\right)$$

donde t_i es el tiempo observado o de censura para cada individuo i y $\mathbf{x}_i$ es el vector de covariables para cada individuo i, como método para determinar el modelo paramétrico.

6.4 Modelo paramétrico versus modelo de Cox

La escogencia del modelo de análisis de supervivencia más adecuado aún es un problema no resuelto.

Sin embargo, dependiendo de los propósitos de la investigación o el estudio pudiera ser más adecuado un modelo paramétrico o un modelo de Cox.

Si lo que se pretende es comparar riesgos entre distintos niveles de las covariables, que suele suele ser el inteés de los estudios médicos, probablemente sea más adecuado utilizar un modelo de Cox.

Si el interés esta basado en obtener informaciones asociadas con parámetros como medias, varianzas, entre otras, como suele ser el interés en el área de la industría, probablemente se recomiende el uso de un modelo paramétrico.

Nardi y Schemper (2003) plantean una intersante discusión que probalemente ayuden a la escogencia entre un modelo de Cox y un

modelo paramétrico.

6.5 Ejemplo 6.1

En esta sección tomaremos como base el objeto cox1 del ejemplo 5.1.

6.5.1 Cálculo de la función de riesgo

Con la librería (package) survival no es posible determinar la función de riesgo por lo que nos valdremos de un artificio.

Consideremos, la función de riesgo estimada definida mediante:

$$\hat{\lambda}(t) = \frac{\hat{f}(t)}{\hat{S}(t)}$$

La cual puede usarse, sin ningún problema, para el caso del tiempo continuo.

Ahora bien, para el caso discreto, en el tiempo t_i, la función de masa de probabilidad puede estimarse mediante:

$$\hat{f}(t_i) = \hat{S}(t_{i-1}) - \hat{S}(t_i)$$

Y la función de sobrevivencia en el tiempo t_i se estima mediante $\hat{S}(t_i)$.

Por consiguiente, la función de riesgo para el caso discreto, puede estimarse mediante:

$$\hat{\lambda}(t_i) = \frac{\hat{S}(t_{i-1}) - \hat{S}(t_i)}{\hat{S}(t_i)}$$

Que es la función que utilizaremos y que podemos calcular mediante los comandos:

> # Determinacion del modelo parametrico mas adecuado:
> # Aislamiento de la funcion de sobrevivencia del objeto cox1 (surv1):
> surv1< −survfit(cox1)$surv
> # Aislamiento del tiempo del objeto cox1 (time1):
> time1< −survfit(cox1)$time > # Creacion del objeto donde se va a almacenar la funcion de riesgo (haz1):
> haz1< −rep(0,length(surv1))
> # Calculo de la funcion de riesgo:
> for (i in 2:length(surv1)) {haz1[i]< −(surv1[i-1]-surv1[i])/surv1[i]}

Pero en la función de riesgo anterior sobra el primer valor, que no ha sido calculado y, al eliminar este valor, también hay que eliminar el primer valor del tiempo, esto se hace a través de los comandos:

> # Eliminacion del primer termino de haz1
(no calculado y con valor igual a cero en este caso):
> haz1< −haz1[haz1>0]
> # Eliminacion del primer termino de surv1
(en este caso el valor es igual a cero, hay que verificar antes):
> time1< −time1[time1>0]

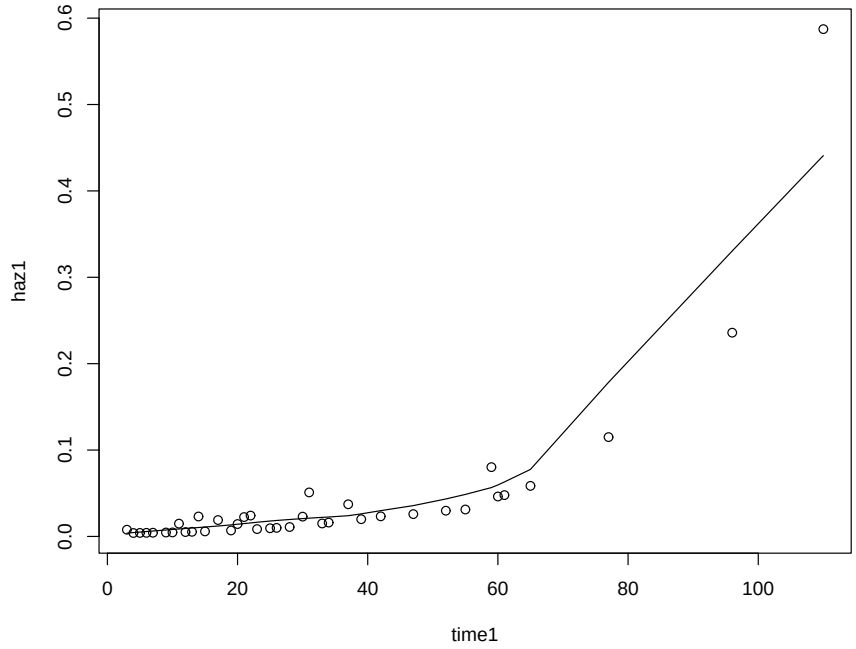
6.5.2 Gráfico para identificar el mejor modelo paramétrico

La determinación del mejor modelo paramétrico lo podemos hacer graficando la función de riesgo estimada versus el tiempo conjuntamente con la curva suavizada, esto lo podemos hacer con los comandos:

> # Grafico (ploteo) de haz1 versus time1:
> plot(time1,haz1,main="Gráfico No. 13. Ploteo de haz1

versus time1 con curva suavizada")
> # Curva suavizada del grafico de haz1 versus time1:
> lines(lowess(time1,haz1,iter=0))

Gráfico No. 13. Ploteo de haz1 versus time1 con curva suavizada



Y observando el Gráfico No. 13 y comparándolo con la galería de funciones de riesgos de los modelos paramétricos teóricos (Meeker y Escobar, 1998), se pudiera pensar que el mejor modelo paramétrico es el Gaussiano o Normal, que es el modelo que utilizaremos en la siguiente sección.

6.5.3 Ajuste del modelo paramétrico

El modelo paramétrico más adecuado (Gaussiano o Normal), lo ajustamos utilizando el comando:

```
> survreg1< −survreg(Surv(meses,censor2)∼diabetes + edad
+ quetelet, data=dpa, na.action=na.exclude, dist="gaussian")
```

Y con el comando survreg1 (o print(survreg1)) se obtiene la siguiente salida:

```
Call:
survreg(formula = Surv(meses, censor2) ∼ diabetes + edad + quetelet,
data = dpa, na.action = na.exclude, dist = "gaussian")

Coefficients:
 (Intercept)        diabetes          edad      quetelet
 42.0167886   -14.3060760   -0.6283562   2.0560862

Scale= 32.4998

Loglik(model)= -339.4 Loglik(intercept only)= -347.7
Chisq= 16.63 on 3 degrees of freedom, p= 0.00084
n=233 (13 observations deleted due to missing)
```

De la salida anterior podemos ontener los coeficientes estimados para el modelo paramétrico, el parámetro de escala y la significación del modelo, es este caso concluimos que el modelo Gaussiano es significativo (al 10%).

A través del comando summary(survreg1) se obtiene la salida:

```
Call:
survreg(formula = Surv(meses, censor2) ∼ diabetes + edad + quetelet,
data = dpa, na.action = na.exclude, dist = "gaussian")
```

	Value	Std. Error	z	p
(Intercept)	42.017	18.5536	2.26	0.02354
diabetes	-14.306	8.0440	-1.78	0.07533
edad	-0.628	0.2204	-2.85	0.00436
quetelet	2.056	0.8345	2.46	0.01374
Log(scale)	3.481	0.0888	39.22	0.00000

Scale= 32.5

Gaussian distribution
Loglik(model)= -339.4 Loglik(intercept only)= -347.7
Chisq= 16.63 on 3 degrees of freedom, p= 0.00084
Number of Newton-Raphson Iterations: 4
n=233 (13 observations deleted due to missing)

De donde obtenemos la información para determinar si cada una de las covariables es significativa en el modelo. En este caso concluimos que diabetes, edad e índice de Quetelet son significativos para un nivel del 10%.

Utilizando el comando names(survreg1) pueden observarse todos los componentes del objeto survreg1.

Adicionalmente, podríamos hacer predicciones o análisis de residuos, pero esto no será tratado en este curso.

Referencias

[1] Allison, P.D. (1995). *Survival Analysis Using the SAS® System: A Practical Guide*. Cary, NC.: SAS Institute, Inc.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. y Keiding, N. (1993). *Statistical Models Based on Counting Processes*. N.Y.: Springer-Verlag.

[3] Borges, R. (2002). *Análisis de Supervivencia Aplicado a un Caso de Diálisis Renal: Diálisis Peritoneal en el Hospital Clínico Universitario de Caracas y Hemodiálisis en el Hospital de Clínicas Caracas, 1980-2000*. Tesis para obtener el grado de M. Sc. En Estadstica Aplicada, Mérida: Instituto de Estadística Aplicada y Computación, ULA. (Disponible en http://tesis.saber.ula.ve/theses/available/etd-06202003-171618/).

[4] Collett, D. (2003). *Modelling Survival Data in Medical Research, 2da. Edicíon*. Boca Ratón: Chapman & Hall.

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

[6] Fleming, T.R. y Harrington, D.P. (1984). Nonparametric estimation of the survival distribution in censored data. *Communications in Statistics. Theory and Methods*, **13:** 2469-2486.

[7] Fleming, T.R. y Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. N.Y.: John Wiley & Sons, Inc.

[8] Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, **33:** 1-26, Londres: Her Majesty's Stationery Office.

[9] Harrington, D.P. y Fleming, T.R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, **69:** 553-566.

[10] Hosmer, D.W. y Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. N.Y.: John Wiley & Sons, Inc.

[11] Insightful Corporation (2001). *S-PLUS 6 for Windows Guide to Statistics, Volume 2*. Seattle, WA: Insightful Corporation, Inc.

[12] Johnson, N.L., Kotz, S. y Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1, 2da Edición*. N.Y.: John Wiley & Sons, Inc.

[13] Johnson, N.L., Kotz, S. y Balakrishnan, N. (1995). *Continuous Univariate Distributions, Volume 2, 2da Edición*. N.Y.: John Wiley & Sons, Inc.

[14] Kalbfleisch, J.D. y Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data, 2da Edición*. N.Y.: John Wiley & Sons, Inc.

[15] Kaplan, E.L. y Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53:** 457-481.

[16] Klein, J.P. y Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. N.Y.: Springer-Verlag.

[17] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data, 2da Edición*. N.Y.: John Wiley & Sons, Inc.

[18] Meeker, W.Q. y Escobar, L.A (1998). *Statistical Methods for Reliability Data*. N.Y.: John Wiley & Sons, Inc.

[19] Miller, R.G. (1981). *Survival Analysis*. N.Y.: John Wiley & Sons, Inc.

[20] Nardi, A. y Schemper, M (2003). Comparing Cox and parametric models in clinical studies. *Statistics in Medicine*, **22:**3597-3610.

[21] Nelson, W.(1969).Hazard plotting for incomplete failure data. *Journal of Quality Technology*, **1:** 27-52.

[22] R Development Core Team (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

[23] S original by Terry Therneau and ported by Thomas Lumley (2005). *survival: Survival analysis, including penalised likelihood*. (R package version 2.17)

[24] Therneau, T.M. y Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. N.Y.: Springer-Verlag.

[25] Therneau, T.M., Grambsch, P.M. y Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, **77:** 147-160.