Capítulo 3

Análisis de duración

"*El futuro está al alcance de cualquiera a una* tasa *de sesenta minutos a la hora*"
C. S. Lewis

En este capítulo tenemos como objetivo hacer una revisión de las principales características de los modelos de duración. En la primera sección se discuten las peculiaridades más relevantes de los datos de duración que explican la necesidad de utilizar técnicas específicas para analizar este tipo de datos. En las secciones dos y tres se abordan los modelos de duración más sencillos. En concreto, en la sección segunda se tratan aquellos modelos que se construyen partiendo de la consideración de que la variable duración es continua mientras que en la sección tercera se hace un repaso de los modelos de duración discretos. A continuación, en las dos siguientes secciones se consideran modelos de duración -continuos o discretos- más complejos que estudian aspectos como la inclusión de variables dependientes del tiempo y la existencia de heterogeneidad no observada.

3.1 Conceptos fundamentales

3.1.1 Introducción

El análisis de datos referidos a la duración de un suceso económico ha experimentado un gran desarrollo durante los últimos veinte y cinco años, en concreto, a partir de los artículos de Lancaster (1979) y Nickell (1979), pioneros en la aplicación del *análisis de supervivencia* en este campo científico. El análisis de supervivencia se desarrolló inicialmente para estudiar la mortalidad de una población, de ahí su nombre, pero con el tiempo -y tras ciertas adaptaciones- se ha ido aplicando a otros campos de la ciencia donde ha sido rebautizado con diferentes acepciones como *análisis de fiabilidad* en ingeniería, *análisis histórico de eventos* en sociología o *análisis de duración* en economía.

En el campo económico se ha utilizado para investigar fenómenos tan diversos como el tiempo que transcurre hasta que un negocio quiebra, la duración de la situación de pobreza de los hogares o el tiempo que se mantiene en el poder un partido político entre otros muchos ejemplos. Dentro del ámbito del mercado de trabajo, se ha aplicado fundamentalmente para analizar la duración del desempleo y, más concretamente, para detectar su posible relación con la percepción de una prestación (Atkinson y Micklewright, 1991).

Como queda reflejado en los ejemplos anteriores, en este tipo de estudios la variable objeto de análisis es la duración de un suceso[1]. Esta variable mide la longitud del tiempo que un individuo pasa en una determinada situación o estado (estado inicial o de origen), o, lo que es lo mismo, el tiempo que transcurre hasta que el individuo *transita* hacia otro estado (estado final o de destino)[2]. El paso de un estado a otro vendrá determinado por la ocurrencia de un suceso[3]. En nuestro caso, la variable duración recoge el tiempo que un individuo tarda en encontrar un empleo significativo tras salir por primera vez

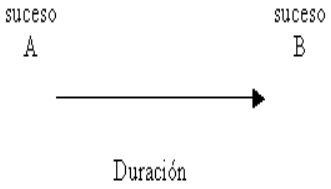
[1] Otros términos con los que se acostumbra a designar la duración son el tiempo de fallo o el tiempo de supervivencia. Estas expresiones resultan más acertadas para recoger el tipo de suceso que se analiza en otras ciencias -el tiempo que tarda en fallar un componente electrónico en el caso de la ingeniería o el tiempo de supervivencia a una enfermedad en el caso de la medicina- aunque aparecen con cierta asiduidad en algunos textos económicos.

[2] Los estados posibles han de ser exhaustivos y mutuamente excluyentes. Además, en general, para concretar el estado de destino suele recurrirse a una definición negativa en el sentido de que éste incluye todas aquellas situaciones distintas a la que define el estado de origen.

[3] Este suceso no tiene porque ser único. Así, por ejemplo, en el caso de la duración del desempleo, la salida del paro puede deberse a que el individuo ha encontrado un empleo o a que ha pasado a la inactividad

del sistema educativo. Esto es, el estado de origen es ese periodo intermedio en el que el individuo aún no ha conseguido establecerse de forma definitiva dentro del mercado de trabajo. Y el estado de destino se alcanza cuando el individuo acepta una oferta de empleo con las características que hemos considerado que debe tener un empleo significativo.

Figura 3.1: **Esquema de una transición**



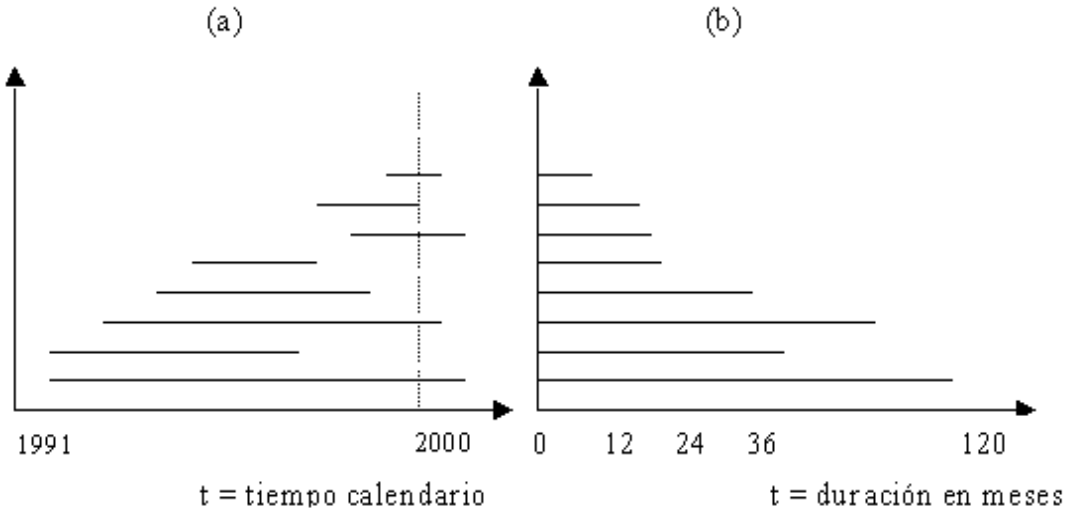
Por lo tanto, para la correcta valoración de la duración de un suceso se requiere establecer un *origen* en el tiempo, una definición precisa de cuando finaliza y una *escala* de medición del tiempo. Estos tres elementos, que han de estar bien ensamblados, son las piezas claves sin las cuales el análisis de datos de duración no funcionaría correctamente.

El origen del tiempo es ese momento a partir del cual cada individuo empieza a estar sujeto al riesgo de que acontezca el suceso (o sucesos) que provocaría el cambio de estado. Este origen no tiene porque coincidir con la fecha en la que el individuo empieza a ser observado y puede venir definido de diferentes modos. Así, como señala Allison (1995), el origen puede ser una edad -por ejemplo, la edad legal para comenzar a trabajar-, un momento en el que ocurre un suceso -el instante en el que se sale del sistema educativo- o una fecha -aquella en la que se empieza a buscar un empleo. Se trata, por lo tanto, de fijar con precisión y claridad en que momento ha de ponerse en marcha el cronómetro que nos va a medir el tiempo, de tal manera, que la especificación de un origen común para todos los individuos nos permita hacer comparaciones entre los mismos al ser sus duraciones equiparables (Cox y Oakes, 1984).

En la figura 3.2 (a), a modo ilustrativo, se muestran algunas de las duraciones del proceso de transición entre el sistema educativo y el mercado de trabajo. Cada línea horizontal representa la duración correspondiente a un individuo y la línea vertical con trazo discontinuo recoge el último periodo en el que se observó cual era su situación respecto al suceso analizado. Como se aprecia en el gráfico, la fecha de inicio no es común para todos pero el hecho de que su origen en el tiempo sea el mismo, nos permite

construir la figura 3.2 (b). En ella, se han trasladado todas las duraciones a ese punto imaginario en el tiempo en el que todos los individuos se encuentran en las mismas condiciones de partida y, por tanto, sus duraciones pueden compararse.

Figura 3.2: **Representación gráfica de las duraciones generadas a partir del módulo de transición del sistema educativo al mercado de trabajo**



De igual modo, el instante que marca el final de la duración, y que nos indica cuando hemos de detener el cronómetro, debe estar bien definido. Dado que los dos momentos que acotan la duración (el inicio y fin de la misma) vienen determinados por los estados de origen y destino, basta con precisar claramente el contenido de esos estados, y en particular, establecer cuáles son los sucesos que dan lugar a esos estados para delimitar la duración. En nuestro caso, significaría concretar qué se entiende por salida del sistema educativo y por consecución de un primer empleo significativo. En el capítulo 4 fijaremos con mayor detalle la dimensión de esos dos estados.

El tercer elemento a tener en cuenta es la escala del tiempo. En este sentido, el criterio básico consiste en elegir la unidad más pequeña relevante para el suceso que se está estudiando. Por ejemplo, a la hora de medir el tiempo que transcurre hasta que se accede a un primer empleo, podemos considerar el número de días o de meses que pasan antes de conseguir ese empleo. En principio, dado que el suceso puede ocurrir cualquier día desde que se sale del sistema educativo, tendría más sentido utilizar como unidad de medida el día. Ahora bien, las herramientas o las fuentes estadísticas disponibles

normalmente van a ser limitadas, lo que nos va a obligar en ocasiones a trabajar con unidades de medida mayores a las ideales. Una consecuencia de la pérdida de precisión en la medición del tiempo es la traslación desde el campo continuo al campo discreto de la variable duración. La distinción entre datos de duración en tiempo continuo o en tiempo discreto tiene importantes implicaciones ya que la mayor parte del análisis estadístico (definición de las funciones, construcción y estimación de los modelos, etc.) está condicionado por esta cuestión.

La naturaleza de la variable duración no es por sí misma una condición suficiente para plantear una nueva metodología alternativa a la tradicional. Dos aspectos característicos de los datos de duración, hasta ahora no puestos de relieve, hacen que tanto el análisis de regresión como los modelos de elección discreta no sean los procedimientos más adecuados para el tratamiento de este tipo de variable. En concreto, nos estamos refiriendo a la existencia de datos censurados y a la presencia de variables explicativas que varían con el tiempo.

En primer lugar, la *censura* es algo inherente a los datos de duración. Al tratarse de una variable que va evolucionando a medida que transcurre el tiempo, éste es un factor clave para recoger información sobre la duración. Por ello, es habitual detectar individuos que aún no han experimentado el suceso una vez ha finalizado el periodo de observación. Las duraciones asignadas a estos individuos son denominadas censuradas ya que no se conoce su valor exacto. Por otro lado, el hecho de que la duración se desarrolle en el tiempo supone que algunos de los factores que pueden tener influencia sobre la misma no se mantienen en sus valores iniciales durante todo el periodo de observación. Estas dos cuestiones serán tratadas de un modo más extenso a continuación.

Resumiendo, en este apartado hemos perfilado algunas de las principales características de los datos de duración que exigen la utilización de técnicas más específicas, contenidas en lo que se conoce como *análisis de duración*. Básicamente, la duración es una variable aleatoria no negativa que, en general, se verá afectada por la censura y que vendrá determinada por factores que, a su vez, pueden variar con el tiempo. El objeto de este estudio va a ser la ley de probabilidad que rige la duración y la modelización se va a llevar a cabo incorporando aquellos factores que influyen de forma directa en la duración. Para ello contamos con el análisis de duración que incluye una colección de procedimientos estadísticos adecuados para modelizar el tiempo que transcurre hasta que un suceso tiene lugar. A continuación, en este capítulo, vamos a sintetizar los aspectos fundamentales de este tipo de procedimientos.

3.1.2 Censura

En general, una observación se dice que está censurada cuando no se conoce con exactitud el tiempo que ha transcurrido entre los dos sucesos que delimitan la duración. Este desconocimiento puede ser debido a que se ignora el momento en el que se inició la duración, a que no se ha observado el momento en el que finalizó, o a la conjunción de ambas situaciones.

Dada la importancia de la censura en este tipo de datos es lógico plantearse cuál es el tratamiento adecuado de la misma, sobre todo si se tiene en cuenta que no siempre los individuos con duraciones censuradas forman un subconjunto aleatorio de la población sino que, en muchas ocasiones, sus duraciones son las más largas. Por esta razón, algunas prácticas como eliminar las observaciones que presentan censura o proceder como si estas observaciones no estuvieran censuradas, resultan insatisfactorias desde un punto de vista estadístico. Como veremos en el desarrollo de este capítulo, el análisis de duración dispone de las herramientas precisas para poder trabajar con la censura lo que le convierte en el procedimiento más apropiado para analizar estos datos.

El número de observaciones censuradas está fuertemente relacionado con la probabilidad de ocurrencia y el periodo de observación del suceso. Por un lado, cuanto más raro es el suceso menor es su probabilidad de ocurrencia lo que conlleva un mayor número de duraciones censuradas. Por otro lado, el diseño del experimento que genera los datos y, en concreto, el periodo de seguimiento de los individuos, también condiciona el número de observaciones censuradas. Así, cuanto más largo es el periodo de observación, más posibilidades hay de que el individuo experimente el suceso. En cualquier caso, hemos de ser conscientes de que en la mayoría de los casos no basta con ampliar el periodo de observación para eliminar la censura dado que existe la posibilidad de que determinados individuos nunca lleguen a experimentar el suceso.

Además de la finalización del periodo de observación, la pérdida de individuos durante el desarrollo del estudio por motivos diversos -cambio de domicilio, fallecimiento o el propio desgaste- es también otra causa bastante común que genera censura. Algunos de estos motivos pueden estar directamente relacionados con la propia ocurrencia del suceso, como por ejemplo cuando un individuo rehúsa seguir contestando a un cuestionario sobre su situación laboral una vez que ya ha encontrado empleo. En este caso, el mecanismo que produce la censura no es totalmente independiente de la ocurrencia del suceso. En la literatura a este tipo de censura se la denomina *censura*

informativa. Coloquialmente, la censura es informativa cuando lo que provoca que la duración esté censurada muestra una relación directa con el mismo proceso de duración.

A pesar de que la validez del análisis de duración se fundamenta en el cumplimiento de la premisa de que la censura es no informativa, no existe en la literatura ningún test que permita contrastarlo. Algunos planteamientos iniciales, como el diseño del experimento, pueden salvaguardarnos de la aparición de esta censura tan problemática. Así, por ejemplo, un cierto control del investigador sobre la finalización del periodo de seguimiento hace que la censura producida por esta razón no esté vinculada al proceso de duración[4].

El tipo de censura al que hemos aludido hasta ahora se caracteriza por el desconocimiento de la fecha de salida del estado inicial, rasgo que identifica a la censura denominada *censura por la derecha*. Dentro de este tipo de censura se distingue entre datos en los que existe *un único valor de censura*, propio de estudios en los que se sigue a una cohorte hasta una fecha determinada que se corresponde con el único valor de censura observado, y datos con *múltiples valores de censura*, que surgen cuando se van perdiendo unidades de la población antes de finalizar el estudio o cuando la población está formada por varias cohortes[5].

Otros tipos de censura son la *censura por la izquierda* y la *censura por intervalos*. Con respecto a la primera, y dentro de las ciencias sociales, esta censura recoge aquellas situaciones en las que se desconoce el origen en el tiempo de la duración, esto es, la fecha de inicio no ha sido observada. En la censura por intervalos se sabe que el tránsito ha tenido lugar dentro de un intervalo pero se desconoce el momento exacto. Este tipo de censura es más frecuente en aquellos estudios en los que el seguimiento de los individuos se realiza a través de una encuesta de carácter periódico. De tal manera que si entre dos encuestas consecutivas el individuo cambia de estado, no se conoce la fecha en la que realizó la transición y la duración vendrá definida por un intervalo.

Dado que la censura por la derecha es sin lugar a dudas la más frecuente, se han

[4]En el caso de censura informativa, Allison (1995) recomienda la inclusión, como variables explicativas, de aquellos factores que pudieran estar relacionados con el origen de la censura. En este sentido, la censura que se produce debido a que el estudio está referido a cohortes con distintas fechas de entrada, queda controlada si se incluye el tiempo calendario como regresor en el modelo de duración.

[5]En las ciencias experimentales también se distingue entre censura tipo I y tipo II. En esta última, el investigador decide finalizar el experimento cuando un número de individuos prefijado de antemano abandona el estado de origen. Klein y Moeschberger (1997) presentan una exposición más detallada sobre los tipos de censura.

llevado a cabo grandes esfuerzos en el desarrollo de esta metodología para conocer cuáles son sus implicaciones y resolver los inconvenientes que conlleva. De ahí que, en la actualidad, su tratamiento no plantee ningún probema estadístico. Respecto al tratamiento de los otros dos tipos de censura, las complicaciones más importantes -aún no resueltas de forma definitiva en la literatura- surgen al estimar los modelos semiparamétricos.

3.1.3 Variables dependientes del tiempo

La mayor parte de los métodos estadísticos que se incluyen dentro del análisis de duración tratan de explicar el comportamiento de la duración de un suceso a partir de los factores que condicionan su evolución. Estos factores englobados en las variables explicativas pueden ser fijos o dependientes del tiempo, entendiendo por tiempo tanto el tiempo calendario como la duración.

Como ya se ha comentado, la posibilidad de incluir variables que varíen en el tiempo es una de las características específicas de los modelos de duración que les diferencia de los modelos convencionales, siendo la propia duración la causa y el efecto de la existencia de este tipo de variables. Por un lado el hecho que la duración se desarrolle en el tiempo, hace que las condiciones que afectan a la misma puedan también variar. Y, por otro lado, el cambio en los factores explicativos también puede afectar a la propia duración, acortándola o alargándola.

De cara a las repercusiones que tiene la introducción de este tipo de variables en estos modelos, vamos a distinguir entre variables dependientes externas e internas, tal y como proponen inicialmente Kalbfleisch y Prentice (1980). Una variable se define como *externa* cuando sus valores se conocen de antemano (externa definida) o cuando aunque se desconocen, su existencia no depende de la "supervivencia" del individuo ni de sus características específicas (externa secundaria)[6]. Dentro del análisis de la duración del desempleo, las variables edad o prestación por desempleo podrían ser ejemplos de variables dependientes externas definidas, y la tasa de paro regional sería un ejemplo de una variable externa secundaria al tratarse de una variable que no está asociada directamente al individuo sino al entorno en el que se encuentra[7]. Por otro lado, aquellas

[6] Las variables que no varían en el tiempo son variables externas fijas.

[7] La prestación por desempleo es una variable que varía principalmente con la duración del desempleo -aunque indirectamente con el tiempo calendario- mientras que la tasa de paro es una variable que varía básicamente con el tiempo calendario.

variables que están definidas en la medida en que el suceso aún no ha ocurrido, esto es, sus valores se pueden determinar tan sólo mientras se dé la "supervivencia" del individuo, son denominadas variables internas. Un ejemplo de variable interna, en un estudio sobre la duración del desempleo de los jóvenes, podría ser la zona geográfica de residencia siempre que el cambio de residencia esté determinado por un cambio en la situación laboral.

Posteriormente, Lancaster (1990) distingue entre variables exógenas y endógenas, demostrando que la estimación de los modelos de duración con variables dependientes del tiempo no supone ningún problema siempre que éstas sean variables exógenas. Según la definición de exogeneidad de Lancaster -variables cuyo comportamiento es el mismo independientemente de que el individuo esté sujeto a riesgo o no- todas las variables externas son exógenas

A pesar de que la existencia de variables dependientes del tiempo exógenas no representa un obstáculo para la estimación de los modelos de duración, su incorporación requiere una modificación en la estructura de los datos que, dependiendo de la herramienta informática, puede resultar más o menos complicada. No obstante, una vez se han dispuesto los datos de forma correcta, como veremos más adelante, la estimación es relativamente sencilla si bien hay que prestar cierta atención a la hora de interpretar los resultados del modelo ya que puede producirse la ruptura de ciertas hipótesis.

3.1.4 Selección de la muestra

El punto de partida de todo análisis estadístico, una vez fijados los objetivos, es recabar toda la información posible sobre la variable o variables objeto de estudio. En nuestro caso, para construir la variable duración es necesario hacer un seguimiento del individuo a lo largo de un periodo de tiempo más o menos extenso. Atendiendo al tipo de seguimiento, podemos distinguir entre estudios *retrospectivos*, en los que a los individuos seleccionados se les hace recordar cuál era su situación y cómo se vio modificada durante el periodo de seguimiento, o *prospectivos*, en cuyo caso a los individuos seleccionados se les va preguntando periódicamente estas mismas cuestiones. Independientemente de si el estudio tiene caracter retrospectivo o prospectivo, antes de llevar a cabo el seguimiento de la población, normalmente es necesario que se realize algún tipo de selección de los individuos que la conforman.

Dado que el esquema de selección de la muestra tiene importantes implicaciones

en la estimación de estos modelos, produciendo en determinados casos problemas de sesgos en los estimadores, repasamos en esta sección algunos de los esquemas que mayoritariamente se han utilizado en este tipo de modelos. Siguiendo a Lancaster (1990) contemplamos, en términos generales, tres esquemas de selección. En primer lugar, la selección de una muestra de individuos en el momento en que *entran* en el estado de interés -por ejemplo, el desempleo- y el seguimiento de los mismos de forma que se recabe información sobre el tiempo que han estado desempleados y, en ocasiones, sobre cuál es su situación tras salir del desempleo. Este tipo se selección recibe el nombre de *selección flujo* pues la muestra está formada por individuos que fluyen hacia el estado cuya duración se está analizando. En segundo lugar, la selección se puede llevar a cabo a partir de un stock de desempleados, es decir, se selecciona a individuos que ya se encuentran en el paro. En esta *selección stock*, los individuos que están más tiempo desempleados tienen una probabilidad mayor de ser seleccionados lo que da origen al sesgo por selección stock. A los dos esquemas ya expuestos, Lancaster (1990) añade un tercero en el que la selección se lleva a cabo a partir de una población en la que no se impone ningún tipo de condicionante referido al suceso que se pretende analizar.

En relación a esta cuestión, algunos autores consideran que la selección de la muestra puede verse como un truncamiento de los datos[8], distinguiendo entre truncamiento por la izquierda en el que sólo los individuos que han "sobrevivido" al menos un determinado tiempo forman parte de la muestra y, truncamiento por la derecha en el que sólo los individuos que llegada una fecha han experimentado el suceso son incluidos en la muestra. En ambas situaciones se excluye de forma sistemática una parte de la población. El truncamiento por la izquierda se correspondería con lo que Lancaster define como seguimiento de una selección stock de una población.

Como veremos, la estimación de estos modelos de duración por máxima verosimilitud requiere, para que los estimadores obtenidos no presenten sesgos, que la función de verosimilitud sea especificada correctamente según el esquema de selección utilizado.

3.1.5 Clasificación

El análisis de duración forma parte de un conjunto de técnicas estadísticas que genéricamente se ha denominado *análisis de transición*. En este caso, el investigador

[8] Frente a la censura, el truncamiento supone un desconocimiento absoluto de la duración para algunos de los individuos.

dispone de mayor información sobre una etapa más extensa de la vida laboral de cada individuo. En concreto, los datos contienen información sobre los estados -y las duraciones correspondientes a cada estado- por los que los individuos han transitado durante una parte de su trayectoria laboral. En estudios con estas características se pone mayor enfásis en el número y tipo de transición y en el orden en que se realizan las mismas, dado que el interés está en discernir cuáles son los determinantes de esas trayectorias laborales.

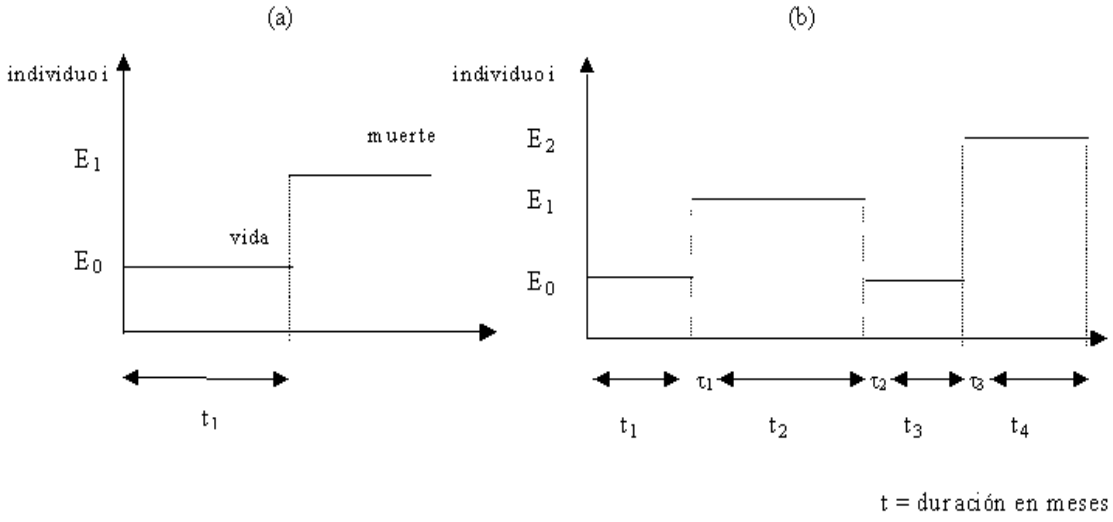
El análisis de transición plantea la utilización de diferentes medotologías según el número de estados posibles -uno o varios- y según el número de periodos observados -uno o varios, incluyendo dentro de esta distinción los modelos de duración como un caso particular. Estrictamente, consideramos que únicamente aquellas situaciones en las que se observa un solo periodo de duración -o lo que es lo mismo, una única transición- son las que se encuadran dentro de los modelos de duración. Ahora bien, atendiendo a esa doble clasificación, y como resultado de la combinación de las cuatro categorías consideradas, se puede distinguir también entre modelos de duración con *riesgos simples*, en los que únicamente se observa un periodo de duración y el estado de destino es único[9], y modelos de duración con *riesgos en competencia* en los que al finalizar la duración el individuo puede salir a varios estados que compiten entre sí. Y, por otro lado, distinguimos frente a los modelos de duración con un único periodo, los llamados modelos con *sucesos recurrentes* en los que el individuo puede volver a transitar al mismo estado con posterioridad.

En nuestro caso, los datos de duración referidos a un individuo i vendrán recogidos en un vector (t_i, δ_i, E_i) cuyas componentes hacen referencia: a la duración observada (t_i), a la condición de la duración en cuanto a la censura (censurada $(\delta_i = 0)$ o no censurada $(\delta_i = 1)$) y al estado ocupado después de la transición (E_i). Inicialmente, y para aquellos modelos que hemos denominado modelos de riesgos simples, el estado final se desconoce o es una información que no es tenida en cuenta.

De forma general, en el análisis de transición, la información correspondiente a un individuo i de la población está contenida en un vector $((t_{1i}, \delta_{1i}, E_{1i}), (t_{2i}, \delta_{2i}, E_{2i}), ..., (t_{li}, E_{li}, \delta_{li}))$ en el que se incluye una trayectoria más amplia del historial del individuo respecto al suceso considerado, en concreto, l duraciones (véase la figura 3.3).

[9] También llamados *modelos de vida-muerte*, término heredado del análisis de supervivencia, aludiendo a los estados inicial y final propios de estos estudios.

Figura 3.3: **Representación gráfica de diferentes modelos de transición: (a) modelo con un único periodo y riesgos simples, (b) modelo con varios periodos y riesgos en competencia**



3.2 Análisis de duración en tiempo continuo

3.2.1 Función de supervivencia y función de riesgo

Sea T la variable aleatoria continua no negativa que recoge la duración del suceso y denotemos con t las realizaciones de dicha variable. La caracterización de una variable, en términos estadísticos, se realiza habitualmente a través de su función de densidad o de su función de distribución. Sin embargo, cuando se trabaja con datos de duración es más normal (y conveniente) utilizar otras dos funciones: la función de supervivencia y la función de riesgo. Este cambio de perspectiva nos va a permitir caracterizar mejor el tipo de datos con los que estamos trabajando, ya que ayuda a interpretar de una forma más natural el proceso que genera las duraciones.

Con el objeto de presentar estas dos nuevas funciones, previamente repasamos de forma breve los conceptos ya conocidos, el de función de densidad y función de distribución, para que el lector logre comprender las diferencias que los separan. La *función de densidad* recoge la probabilidad de que la duración sea igual a t, esto es,

$$f(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t)}{\Delta t}$$

De forma idéntica se podría caracterizar la variable T a partir de su *función de distribución*, que informa sobre la probabilidad de que la duración sea menor o igual a t

$$F(t) = p(T \leq t)$$

Sin embargo, en los modelos de duración el suceso complementario al que recoge la función de distribución resulta más interesante. Por ello se define la *función de supervivencia*, que recoge la probabilidad de que la duración sea mayor a t[10],

$$S(t) = p(T > t)$$

Consecuencia de esta definición es que la función de supervivencia, al igual que la función de distribución, sólo toma valores comprendidos entre 0 y 1. Además, se trata de una función monótona decreciente verificándose que si $t_1 < t_2$ entonces $S(t_1) > S(t_2)$, por lo que partiendo del valor uno va disminuyendo hasta anularse[11]. Lo que interesa del comportamiento de esta función es la velocidad a la que se va aproximando al valor cero, y para ello resulta más cómodo analizar la representación gráfica de esta función, denominada curva de supervivencia. En la práctica, este gráfico suele tomar la forma de una función escalonada y no siempre alcanza su valor mínimo debido a la presencia de censura.

Las funciones definidas hasta ahora son un acercamiento a la caracterización de una variable aleatoria desde un punto de vista incondicional. En el caso que nos atañe, resulta más conveniente razonar en términos de probabilidades condicionadas. Así, definimos la *función de riesgo*[12] -que no es más que una función de densidad condicionada- como:

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t / T \geq t)}{\Delta t} \tag{3.1}$$

[10] Algunos textos definen la función de supervivencia como la probabilidad de que la duración sea mayor o igual a un valor t. En el caso que nos ocupa, modelos de duración en tiempo continuo, ambas definiciones estarían recogiendo lo mismo ya que los sucesos puntuales tienen probabilidad nula.

[11] La función de supervivencia es igual a 1 en el origen del tiempo ($T = 0$) ya que se supone que la probabilidad de que ocurra el suceso es uno ($S(0) = p(T > 0) = 1$). En algunos estudios es posible admitir probabilidades no nulas para $T = 0$, como por ejemplo sería el caso si en un estudio sobre la consecución de un primer empleo, algunos estudiantes encuentran el empleo antes de salir del sistema educativo. Esto dificulta en cierta medida el análisis de estos datos de duración (Mouchart, 1999).

[12] En la literatura, esta función ha recibido varios y diferentes nombres como la inversa del ratio de Mills, el tanto instantáneo de fallecimiento, etc.

esto es, recoge el riesgo instantáneo de que el suceso tenga lugar en t condicionado a que hasta ese momento todavía no había ocurrido.

Dada la trascendencia de esta función en los modelos que exponemos, vamos a abordar el contenido de esta expresión matemática desmenuzando los elementos que aparecen en su expresión para, de esta manera, conseguir comprender su significado. Al suponer que T es una variable continua, la probabilidad condicionada de que acontezca el suceso para cualquier valor de t no habiendo ocurrido todavía, va a ser siempre nula. De ahí que se plantee esta probabilidad condicionada no para un instante de tiempo sino para un intervalo suficientemente pequeño. De esta manera quedaría explicado el contenido del numerador de la función de riesgo,

$$p(t \leq T < t + \Delta t / T \geq t)$$

Por otro lado, la probabilidad así construida depende de la longitud del intervalo de tiempo considerado. Para contrarrestar esta dependencia se divide por Δt de tal forma que la probabilidad ya no depende del tamaño del intervalo. Al dividir se obtiene una probabilidad por unidad de tiempo lo cual, como veremos, tiene serias implicaciones en la interpretación de la función de riesgo pues deja de ser una probabilidad. El último paso que damos para llegar a la expresión que aparece en la ecuación 3.1 es tomar límites, de tal manera que estamos dando a la función de riesgo ese carácter de instantáneo al que hemos hecho referencia en su definición.

Como se ha subrayado, resulta más útil trabajar con funciones que vienen definidas en términos de probabilidades condicionadas, como es el caso de la función de riesgo. Para mostrar la capacidad de esta función a la hora de interpretar fenómenos que definen la duración de un suceso proponemos el siguiente caso. Supongamos una muestra de personas que entran por primera vez al desempleo y a las cuales seguimos durante doce meses. Transcurrido dos meses, por ejemplo, nos preguntaremos cuál es el riesgo de que un individuo cualquiera salga del paro en ese momento. Ahora bien, para que ese individuo esté sujeto al riesgo concreto de abandonar el desempleo en ese momento, es necesario que se haya mantenido desempleado durante los dos primeros meses. De ahí que la pregunta sobre cuál es la probabilidad de abandonar el desempleo en un instante t no habiendo ocurrido el suceso hasta ese momento tenga más sentido en este contexto[13].

[13] Allison (1995) interpreta esta función recurriendo a un símil realmente elocuente en el que la función de riesgo mide la *velocidad* a la que el proceso se va completando una vez se ha alcanzado el momento t.

De interés cuando se trabaja con datos de duración, es el comportamiento de esta función de riesgo con respecto al tiempo, lo que se conoce, dentro de la literatura, con el nombre de *dependencia de la duración*. En relación a esta cuestión, interesa determinar si el mero paso del tiempo tiene cierta influencia sobre la probabilidad de que el suceso tenga lugar. Normalmente, mediante un estudio de la forma de la función de riesgo es posible realizar este análisis de dependencia. En particular, se habla de dependencia positiva si a medida que aumenta la duración, las posibilidades de que ésta finalice son mayores. Por el contrario, la dependencia de la duración será negativa cuando la probabilidad de que finalice se reduce con el tiempo. Esta dependencia de la duración se mide a partir de la derivada de la función de riesgo con respecto a t. En concreto, si $\partial h(t)/\partial t > 0$ hablamos de dependencia positiva y si $\partial h(t)/\partial t < 0$ de dependencia negativa. En el mundo real, la función de riesgo no tiene un comportamiento monótono (creciente o decreciente) como el descrito, sino que la forma de la función de riesgo es tan variada como fenómenos se pueden estudiar. Uno de los objetivos de cualquier investigación que tenga como variable dependiente la duración será el de examinar si existe dependencia y, en su caso, identificar en qué sentido se da.

Las dos funciones que tienen una mayor papel en el análisis de duración, la función de riesgo y la función de supervivencia, se mueven en sentido contrario. Esto es, cuanto mayor es la probabilidad de que el suceso tenga lugar, menor es la probabilidad de que se sobreviva. Por lo tanto, si en el gráfico de la función de riesgo observamos un crecimiento rápido de la función, en la curva de supervivencia este crecimiento se traduce en una pendiente decreciente más acusada.

Por último, también definimos la *función de riesgo acumulada*

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

que se puede interpretar como la suma de riesgos a los que se enfrenta un individuo hasta t. Como veremos más adelante, va a resultar ser una función de gran utilidad a la hora de evaluar los modelos.

Aunque la utilización de la función de riesgo permite un análisis de los datos de duración que resulta teórica e intuitivamente más atractivo, hemos de ser conscientes

Por ejemplo, en el supuesto de que $h(t)$ sea constante en el tiempo, esto es, la velocidad se mantiene fija hasta el final, un valor de 0,15 en la función de riesgo de que un individuo esté desempleado, siendo la unidad de medida del tiempo el mes, se interpretaría como el número esperado de veces que el individuo saldría del desempleo en un mes.

de que se trata tan sólo de otro modo de ver las cosas y que cualquiera de las funciones comentadas nos permite caracterizar el comportamiento de esta variable, como se deduce de la existencia de una total correspondencia entre todas estas funciones (véase tabla 3.1).

Tabla 3.1: **Relaciones entre las funciones que caracterizan la variable duración en el caso continuo**

	$f(t)$	$F(t)$	$S(t)$	$h(t)$
$f(t)$	-	$F'(t)$	$-S'(t)$	$h(t) \cdot e^{-\int_0^t h(u)du}$
$F(t)$	$\int_0^t f(u)du$	-	$1 - S(t)$	$1 - e^{-\int_0^t h(u)du}$
$S(t)$	$\int_t^\infty f(u)du$	$1 - F(t)$	-	$e^{-\int_0^t h(u)du}$
$h(t)$	$\frac{f(t)}{1-\int_0^t f(u)du}$	$\frac{F'(t)}{1-F(t)}$	$-\frac{S'(t)}{S(t)} = -\frac{d}{dt}\ln S(t)$	-
$H(t)$	$-\ln(\int_t^\infty f(u)du)$	$-\ln(1 - F(t))$	$-\ln S(t)$	$\int_0^t h(u)du$

3.2.2 Modelización y estimación

En este apartado examinaremos diferentes propuestas para estimar la distribución de probabilidad de la variable T. En concreto, vamos a distinguir entre métodos de estimación no paramétricos y métodos paramétricos o semiparamétricos. El análisis no paramétrico sigue la filosofía de dejar a los datos que hablen por si mismos sin imponer ninguna forma funcional y sin modelizar los efectos de las variables explicativas. Por el contrario, en la estimación paramétrica o semiparamétrica, se parte de un modelo en el que se trata de explicar la duración en función de una selección de variables explicativas, especificando de forma paramétrica, bien en su totalidad o parcialmente, dicha función.

3.2.2.1 Estimación no paramétrica

Hasta mediados del siglo XX, el análisis de supervivencia consistía básicamente en el cálculo y la comparación de las funciones de supervivencia. En la actualidad, la estimación de estas funciones sigue siendo la etapa inicial de cualquier análisis de datos de duración, pero no como fin último sino como un paso intermedio que proporciona un examen preliminar del comportamiento de los datos.

En ausencia de censura, la función de supervivencia se puede estimar a partir de la función de supervivencia empírica, que no es más que la proporción de individuos con una duración superior a t. Dado que como se ha subrayado con anterioridad, la censura es una característica típica de los datos de duración, es necesario disponer de algún método de estimación que tenga en cuenta la información de las observaciones censuradas[14]. Este estimador es el *estimador de Kaplan-Meier* (1958) o estimador del producto-límite[15]. A continuación presentamos los elementos claves para el cálculo de este estimador, resaltando las ideas básicas en las que se inspira.

Sean $t_1, t_2, ...t_N$ las duraciones correspondientes a los N individuos independientes pertenecientes a una población homogénea. En primer lugar, se ordenan de menor a mayor los r valores observados *no censurados*, $t_{(1)} < t_{(2)} < ... < t_{(j)} < ... < t_{(r)}$, donde, en general, $r < N$ debido a que habrá valores observados que se repiten y valores censurados. Para cada una de estas duraciones no censuradas se obtiene,

d_j: el número de individuos cuya duración es igual a $t_{(j)}$,

m_j: el número de individuos cuya duración censurada se encuentra entre $t_{(j)}$ y $t_{(j+1)}$, esto es, aquellos individuos a los que se ha seguido hasta $t_{(j)}$ sin que experimenten el suceso y se les ha perdido antes de comprobar que en $t_{(j+1)}$ siguen estando en el estado inicial, y

n_j: el número de individuos *sujetos a riesgo* definido como $n_j = \sum_{j \geq i}(m_j + d_j)$.

Una vez calculadas las expresiones anteriores, se construye el estimador de Kaplan-Meier como,

$$\widehat{S}(t) = \prod_{t_{(j)} \leq t} \frac{n_j - d_j}{n_j} = \prod_{t_{(j)} \leq t} 1 - \frac{d_j}{n_j}$$

Tal como se deriva de la expresión del estimador, los individuos con observaciones censuradas contribuyen en la construcción del estimador ya que serán individuos sujetos

[14] En este apartado solo consideramos el supuesto de censura por la derecha. En Cleves, Gould y Gutierrez (2002) se discuten otros tipos de censura.

[15] Esta denominación tiene que ver con la forma en cómo se construye a partir del producto de una secuencia de probabilidades condicionadas de supervivencia. Una explicación detallada sobre la misma aparece en Hosmer y Lemeshow (1999).

a riesgo hasta el momento en el que se les deja de observar[16]. La curva de supervivencia, representación de esta función, presentará la forma característica de una escalera ya que se supone que la función se mantiene constante entre dos valores observados no censurados consecutivos.

Como estimador de la varianza del estimador de Kaplan-Meier suele utilizarse el propuesto por Greenwood (1926) conocido como *fórmula de Greenwood*,

$$\widehat{\text{var}}\ (\widehat{S}(t)) = \widehat{S}^2(t) \sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

Por otra parte, a partir del estimador de Kaplan-Meier se puede obtener fácilmente un estimador de la función de riesgo acumulada, teniendo en cuenta la relación existente entre dichas funciones (véase tabla 3.1)[17],

$$\widehat{H}(t_i) = -\ln \widehat{S}(t_i)$$

Sin embargo, no es posible obtener directamente un estimador de la función de riesgo si bien puede aproximarse aplicando algún método de suavizado a la función de riesgo acumulada como señalan Klein y Moeschberger (1997).

Comparación de las curvas de supervivencia: La forma más simple de comparar las funciones de supervivencia correspondientes a dos poblaciones es a través de un gráfico en el que se muestran sus respectivas curvas. Ahora bien, existe en la literatura un amplio número de contrastes que permiten evaluar numéricamente si las curvas de supervivencia son estadísticamente iguales. Lee (1992) enumera y explica varios de ellos como son el test de logaritmo del rango (Mantel y Haenszel, 1959), el test de Wilcoxon (Gehan, 1965 y Breslow, 1970), el test de Tarone-Ware (1977) y el test de Peto (Peto y Peto, 1972). Es necesario resaltar que todos estos contrastes analizan si las funciones de supervivencia son iguales de forma global y no puntualmente para determinados valores de t o intervalos de esta variable. En general, estos tests funcionan comparando el número de "fallos" esperado frente al observado, diferenciándose entre sí en el peso que asignan a cada observación.

[16] Cuando para un valor de t se produce de forma simultánea una observación censurada y una no censurada, por convención se considera que la observación censurada ocurre con posterioridad a la no censurada.

[17] Existe otro estimador de esta función que tiene mejores propiedades muestrales, el *estimador de Nelson-Aalen*, cuya expresión es $\widehat{H}(t) = \sum_{j=1}^{i} \frac{d_j}{n_j}$.

En concreto, el test del logaritmo del rango, en el caso en el que se contrasta la igualdad de las curvas de supervivencia de dos grupos, tiene como numerador,

$$\sum_{j=1}^{r}(d_{1j} - n_{1j}\frac{d_j}{n_j})$$

donde el sumatorio incluye todos aquellos valores de t que son comunes a ambos grupos y $n_{1j}\frac{d_j}{n_j}$ recoge el número esperado de fallos para uno de los grupos bajo la hipótesis de igualdad de las curvas. Dividiendo la expresión anterior por un estimador de su varianza se obtiene el estadístico de este contraste que se distribuye como una χ^2 con un grado de libertad[18]. Este test es bastante potente cuando las funciones cumplen la hipótesis de proporcionalidad que explicaremos más adelante.

Si en la expresión anterior se fija un peso distinto a cada una de las diferencias entre los valores observados y los esperados, entonces se obtienen -dependiendo del peso elegido- los otros estadísticos nombrados anteriormente. Así, por ejemplo, el test de Wilcoxon pondera esas diferencias por el número de individuos sujetos a riesgo en cada t, esto es, su numerador es igual a

$$\sum_{j=1}^{r} n_j(d_{1j} - n_{1j}\frac{d_j}{n_j})$$

Por lo tanto, la utilización de este estadístico supondría dar una mayor importancia a las diferencias correspondientes a aquellos valores de la duración más pequeños, que es cuando el número de individuos sujetos a riesgo suele ser más alto. Por su parte, el test del log-rango, al no considerar ningún tipo de ponderación, da la misma importancia a todas las diferencias observadas.

3.2.2.2 Modelización paramétrica

La estimación paramétrica parte de la idea de que se conocen los principales rasgos de la forma de la función de riesgo y que pueden aproximarse por alguna distribución conocida. Bajo esta premisa, la estimación paramétrica consiste en seleccionar una o varias distribuciones entre aquellas correspondientes a una variable aleatoria continua no negativa, estimar el modelo parametrizado y elegir aquél que sea considerado el mejor según uno o varios criterios de selección. Algunas distribuciones, que han

[18] En el caso más general en el que se comparan k curvas de supervivencia, los grados de libertad de la χ^2 serán $k-1$.

sido frecuentemente utilizadas en la literatura aplicada, son: la exponencial, Weibull, Gompertz, gamma, logarítmico normal, logarítmico logística, Pareto o Gaussiana inversa.

Además, dado que en la mayoría de las ocasiones la población de estudio no puede considerarse homogénea, es habitual tener en cuenta las diferencias existentes entre los individuos de la población mediante la inclusión de determinadas características como factores explicativos de la duración. La incorporación de estas variables explicativas en la especificación de la función de riesgo puede tomar diversas formas, siendo las que más se han popularizado dentro de la aplicación de estos modelos en el campo económico, el modelo de *riesgos proporcionales* (PH) y el modelo de *riesgos acelerados* (AFT).

Modelo de riesgos proporcionales: El modelo de riesgos proporcionales (*proportional hazard model*) se caracteriza porque las variables explicativas actúan de forma multiplicativa, lo que viene a significar que la función de riesgo se obtiene al multiplicar una función de riesgo denominada "base" ($h_0(t)$) por una función no negativa de las variables explicativas ($\varphi(x)$), esto es,

$$h(t;x) = h_0(t)\varphi(x)$$

La forma funcional φ más comúnmente utilizada en las aplicaciones ha sido $e^{x\beta}$, dado que de esta manera se consigue que la función de riesgo resultante sea siempre positiva sin imponer ningún tipo de restricción a los parámetros β. En base a esta expresión de la función de riesgo y teniendo en cuenta las relaciones existentes que presentamos en la tabla 3.1 se derivan sin dificultad el resto de funciones que caracterizan a la variable T (véase tabla 3.2).

Tabla 3.2: **Deducción de la expresión de las funciones que caracterizan la variable duración en los modelos de riesgos proporcionales y en los de riesgos acelerados**

	Modelo de riesgos proporcionales (PH)	Modelos de riesgos acelerados (AFT)
$h(t;x)$	$h_0(t)\varphi(x)$	$h_0(t\varphi(x))\varphi(x)$
$S(t;x)$	$[S_0(t)]^{\varphi(x)}$	$S_0(t\varphi(x))$
$f(t;x)$	$f_0(t)\varphi(x)[S_0(t)]^{\varphi}(x)$	$f_0(t)\varphi(x)$
$H(t;x)$	$H_0(t)\varphi(x)$	$H_0(t\varphi(x))$

De cara a descifrar el alcance de estos modelos de riesgos proporcionales vamos a suponer que el vector de variables explicativas sólo contiene una variable dicotómica que recoge, por ejemplo, el sexo (0 = hombre y 1 = mujer). De este modo, consideramos el grupo de los hombres como el grupo de referencia con respecto al que vamos a comparar el riesgo al que se enfrenta el otro grupo, el de las mujeres. Con este planteamiento la función de riesgo para los hombres es

$$h(t; x = 0) = h_0(t)e^{x\beta} = h_0(t)$$

expresión que nos permite explicar el nombre dado a $h_0(t)$ (función de riesgo base) ya que recoge el riesgo del individuo de referencia para el cual $x = 0$. La función de riesgo del grupo de las mujeres será,

$$h(t; x = 1) = h_0(t)e^{\beta}$$

que comparada con la función de riesgo de los hombres supone un desplazamiento vertical de esta función en una cantidad constante e^{β} para cualquier valor de t.

De este resultado se deriva la principal característica de este modelo, y a la cual debe su nombre, la proporcionalidad. Dados dos individuos con características x_1 y x_2 respectivamente, se verifica que sus funciones de riesgo son proporcionales como se aprecia en la siguiente expresión,

$$\frac{h(t; x_1)}{h(t; x_2)} = e^{(x_1 - x_2)\beta} = cte$$

en la que el término $h_0(t)$ ha desaparecido lo que significa que la relación se mantiene constante en el tiempo. Si retomamos el ejemplo anterior en el que únicamente había una variable explicativa, esta cualidad del modelo supone que la forma de las funciones de riesgo de los dos grupos -hombres y mujeres- sea idéntica y, por tanto, que el efecto de la variable sea tan sólo desplazar verticalmente una de ellas por encima o por debajo de la otra.

Kiefer (1988) demuestra como el modelo de riesgos proporcionales puede ser interpretado como un modelo de regresión lineal con ciertas limitaciones. En concreto, la expresión a la que llega es,

$$\ln H_0(t) = -x\beta - \varepsilon$$

donde la perturbación sigue una distribución específica, la distribución del valor extremo.

Modelo de riesgos acelerados: Si bien es cierto que los modelos de riesgos proporcionales se han extendido muy rápidamente dento del análisis de duración, principalmente en determinados campos científicos, existen otros modelos que han adquirido igualmente cierta importancia como son los modelos de riesgos acelerados (*accelerated failure time models*). En aquellas circunstancias en las que la proporcionalidad que caracteriza al modelo de riesgos proporcionales no es razonable, los modelos de riesgos acelerados pueden resultar realmente útiles.

Bajo el supuesto de riesgos acelerados, las variables explicativas tienen el efecto de cambiar la escala del tiempo[19], esto es, la duración "base" tras la incorporación de los factores explicativos se va a ver transformada tal como se muestra en la siguiente expresión:

$$T = \frac{T_0}{\varphi(x)} \tag{3.2}$$

donde T es la duración reescalada y $\varphi(x)$ es una función de las variables explicativas que se constituye en el factor de aceleración o deceleración del tiempo según sea su signo. En concreto, si $\varphi(x) > 1$ el tiempo se acelera y por lo tanto la duración se verá reducida, y si $\varphi(x) > 1$ el tiempo trancurre más lentamente y la duración se alargará.

La implicación de este supuesto cuando se trabaja con la función de supervivencia es que las variables explicativas tienen como efecto la necesidad de evaluar dicha función en un tiempo reescalado por el factor de aceleración. Así es, partiendo de la función de supervivencia base, en la que no se tiene en cuenta la heterogeneidad de la población, $S_0(t)$, la incorporación de las variables explicativas nos conduce a la siguiente expresión,

$$S(t; x) = S_0(t\varphi(x))$$

Al igual que el modelo de riesgos proporcionales, el modelo de riesgos acelerados también puede expresarse como un modelo de regresión. Si retomamos la ecuación 3.2 para un valor dado de T y tomamos logaritmos en ambos lados de la ecuación, se obtiene,

$$\ln t = -\ln(\varphi(x)) + \ln t_0$$

expresión en la que dependiendo de la distribución de $\ln t_0$ -término que asimilaremos al término de perturbación del modelo- obtendremos un modelo de regresión concreto. En

[19] A diferencia de los modelos de riesgos proporcionales en los que las variables explicativas tenían por efecto reescalar la función de riesgo.

la tabla 3.3 aparecen aquellas distribuciones del término de perturbación (ϵ) que han sido utilizadas en la literatura y las distribuciones de T que llevan asociadas. Como veremos en la parte expositiva de las características de algunos de los modelos paramétricos, es habitual designar a estos modelos por la distribución de T que llevan implícitamente asociada y no por la distribución del término de perturbación.

Tabla 3.3: **Correspondencia entre la distribución de ϵ y la distribución de la duración en los modelos de riesgos acelerados**

Distribución de ϵ	Distribución de T
Valor extremo (1 parámetro)	exponencial
Valor extremo (2 parámetros)	Weibull
Normal	logarítmico normal
Logística	logarítmico logística
Logarítmico gamma	gamma

Si se utiliza la forma funcional $e^{-x\beta}$ para recoger el efecto de las variables explicativas, el modelo de regresión queda de la forma,

$$\ln t = x\beta + \epsilon$$

expresión que nos ayuda a interpretar, de forma más sencilla, los coeficientes β en estos modelos de riesgos acelerados.

Por último, la transformación de las funciones más relevantes al incorporar factores explicativos suponiendo riesgos acelerados aparece también en la tabla 3.2.

Como se ha indicado, las dos especificaciones (PH y AFT) tienen su equivalencia dentro de los modelos de regresión lineal, lo que resulta de gran utilidad para interpretar los coeficientes β de una forma similar a la interpretación de los coeficientes de regresión en los modelos lineales y nos sirve igualmente para ilustrar las dificultades comparativas de ambas especificaciones. Así, en el modelo de riesgos proporcionales, el coeficiente recoge el efecto proporcional de la variable x sobre el logaritmo de la función de riesgo ($\beta = \partial \ln h(t)/\partial x$) mientras que en el modelo de riesgos acelerados recoge el efecto de la variable x sobre el logaritmo de t ($\beta = \partial \ln t/\partial x$). En consecuencia, si bien el signo de estos coeficientes será, en cualquier caso, contrario, no es admisible su comparación pues no están midiendo lo mismo[20].

[20]No obstante, algunos modelos paramétricos tienen la propiedad de poder expresarse en términos de

La consideración de una u otra especificación no parece estar determinada por unos criterios estrictamente estadísticos sino más bien por las prácticas habituales dentro de cada campo científico, la disponibilidad de herramientas informáticas y, en último lugar, de la idoneidad de una forma funcional concreta. No obstante, a continuación presentamos algunas consideraciones extraídas de Jenkins (2004), que pueden ayudar a decantarse por un modelo de riesgos proporcionales o por un modelo de riesgos acelerados. En particular, Jenkins destaca las limitaciones propias de cada especificación y la forma en que cada modelo se adapta para encontrar una solución parcial a las mismas. Por un lado, el modelo de riesgos proporcionales es más restrictivo en cuanto a la posibilidad de elegir entre distintas distribuciones para el término de error si bien permite diferentes transformaciones de la variable duración para conseguir la linealidad en las variables explicativas. Por el contrario, el modelo de riesgos acelerados aunque habilita distintas distribuciones para la perturbación, restringe las posibilidades de transformación de la duración[21].

Una vez hemos presentado y comentado las dos fórmulas de uso común para especificar modelos paramétricos, dedicamos a continuación los siguientes apartados a conocer las principales características de algunos de estos modelos. Ello nos va a permitir también mostrar el mecanismo mediante el cual las dos especificaciones ensamblan los dos componentes básicos de los modelos de duración, las variables explicativas y la propia duración.

Algunos modelos paramétricos

Modelo exponencial. El modelo exponencial es el que resulta de suponer que T sigue una distribución exponencial. Esta distribución ha sido elegida en primer lugar por diversos motivos: su simplicidad matemática, su significación histórica y sus señaladas propiedades. En este apartado vamos a describir sus propiedades suponiendo una función exponencial de parámetro λ, con $\lambda > 0$.

Esta distribución se caracteriza por tener una función de riesgo constante, esto es, $h(t) = \lambda$, lo que significa que si unos datos se ajustan a una distribución exponencial no

un modelo PH y de un modelo AFT, existiendo por tanto una relación directa entre los coeficientes de ambos modelos.

[21] Ridder (1990) propuso un modelo que anidaba los dos modelos (PH y AFT) cuyo principal inconveniente es que no permite la inclusión de datos censurados de ahí su poca repercusión.

mostrarán ningún tipo de dependencia de la duración. A esta propiedad se la conoce como falta de memoria o propiedad del no envejecimiento ("*old as good as new*"), lo que da lugar a su fácil manejo desde un punto de vista matemático, pero también reduce las posibilidades de su utilización en situaciones reales.

El modelo de riesgos proporcionales en el que la función de riesgo base sigue una distribución exponencial de parámetro λ viene recogido en la siguiente expresión:

$$h(t;x) = \lambda e^{x\beta}$$

que quedaría transformado en un modelo en el que la función exponencial contiene un término constante que absorbe el parámetro de la distribución exponencial, esto es, $h_0(t) = e^{\beta_0}$.

El modelo exponencial admite también ser interpretado como un modelo de riesgos acelerados,

$$\ln t = x\beta^* + \sigma\epsilon$$

donde la perturbación, como se recoge en la tabla 3.3, sigue una distribución del valor extremo y $\sigma = 1$[22].

En este caso, por tanto, existe una relación fácilmente deducible entre los parámetros de ambas especificaciones ($\beta = -\beta^*$). Resultado lógico ya que mientras que en el modelo de riesgos proporcionales se plantea la estimación en términos de la función de riesgo, en el modelo de riesgos acelerados la variable dependiente es la propia duración (o el logaritmo de ésta), y la relación entre estas dos variables es inversa, esto es, si un suceso presenta un riesgo elevado de que se produzca, la duración de este suceso tenderá a ser corta.

La distribución exponencial juega un papel semejante a la distribución normal en la modelización de modelos lineales puesto que, como demuestra Lancaster (1990), sea cual sea la forma de la función de riesgo de una variable, ésta podrá ser transformada en otra de riesgo constante (esto es, con distribución exponencial) mediante un cambio en la escala del tiempo. Este resultado tiene cierta trascendencia en los métodos propuestos para seleccionar la distribución que mejor se ajusta a unos datos. En particular, un posible procedimiento para verificar si unos datos se adaptan bien a la distribución

[22]También conocida como distribución Gumbel, la distribución del valor extremo es unimodal y con una ligera asimetría a la izquierda, siendo su función de densidad igual a $f(\epsilon) = e^{\epsilon - e^{\epsilon}}$.

exponencial consiste en representar gráficamente el $\ln(S(t))$ frente al tiempo y comprobar si los puntos están próximos a la recta de pendiente λ que pasa por el origen.

Otra de las desventajas de esta distribución es que depende tan sólo de un único parámetro por lo que pequeñas desviaciones, dan lugar a ajustes muy sensibles. Con independencia de estas cuestiones de fondo, el modelo exponencial se ha convertido en un modelo de referencia en muchas de las aplicaciones empíricas, sobre todo en aquellas que buscan determinar si existe o no dependencia de la duración. En las figuras 3.4-3.9 mostramos gráficamente la forma de las funciones de los modelos paramétricos que se han considerado en esta presentación, entre ellos el modelo exponencial en el que el parámetro es igual a 0,5.

Modelo Weibull. Una variable se dice que sigue una distribución Weibull si su función de riesgo tiene la siguiente expresión:

$$h(t) = \lambda p t^{p-1} \qquad \lambda > 0, p > 0$$

Por lo tanto, se trata de una distribución con dos parámetros, un parámetro de escala λ y un parámetro p que determina la forma de la función de riesgo. En concreto, valores de p superiores a la unidad suponen una función de riesgo creciente[23] y valores de p inferiores a la unidad dan lugar a una función de riesgo decreciente. En el caso de que el parámetro de forma sea igual a la unidad, la distribución queda reducida a una exponencial y, por tanto, su función de riesgo es constante. Consecuentemente, estamos ante una distribución bastante más flexible que la exponencial, siendo capaz de recoger diversas situaciones en las que la función de riesgo tiene siempre un comportamiento monótono, creciente o decreciente, como se muestra en la figura 3.5 en la que se ha representado esta función para distintos valores de p (siendo el parámetro de escala siempre igual a 1).

La utilización de la distribución Weibull para recoger el comportamiento de la función de riesgo base de un modelo de riesgos proporcionales nos conduce a la siguiente expresión del modelo:

$$h(t; x) = h_0(t)e^{x\beta} = \lambda p t^{p-1} e^{x\beta}$$

[23] Siendo más correctos, si $1 < p < 2$ la función de riesgo crece a un ritmo decreciente y si $2 < p < \infty$ la función crece pero a un ritmo creciente.

donde el parámetro λ queda absorbido como constante en el exponente de la función exponencial,

$$h(t;x) = t^{p-1}e^{\beta_0+x\beta}$$

Este modelo tiene la virtud de que también se puede interpretar como un modelo de riesgos acelerados -propiedad que tan sólo cumplen la distribución Weibull y la exponencial- en el que la distribución de la perturbación sigue una distribución del valor extremo (véase tabla 3.3),

$$\ln t = x\beta^* + \sigma\epsilon$$

existiendo, por tanto, también una relación entre los parámetros de ambas expresiones del modelo, que vendrá dada por $\beta^* = -\beta/p$.

Asimismo, el método para comprobar si unos datos se ajustan bien a esta distribución, consiste en construir un gráfico en el que se representa el $\ln(-\ln(S(t)))$ frente a $\ln t$ y verificar en que medida los datos están dispuestos próximos a la recta de ordenada $\ln\lambda$ y con pendiente p. Este resultado se deduce a partir de la expresión de la función de riesgo acumulada de la distribución Weibull, $H(t) = \lambda t^p$ y, en concreto, de aplicar neperianos a esta expresión,

$$\ln(H(t)) = \ln\lambda + p\ln t$$

lo que da como resultado una recta.

La flexibilidad y el hecho de que las funciones que caracterizan esta distribución sean bastante tratables han hecho que crezca el uso y la popularidad del modelo Weibull, sobre todo dentro del campo económico para modelizar la duración del desempleo ya que permite recoger la dependencia de la duración positiva que acostumbra a detectarse en este tipo de datos.

Modelo Gompertz. Se dice que una variable sigue una distribución Gompertz (Gompertz, 1825) de parámetros λ y γ si su función de riesgo se ajusta a la siguiente expresión,

$$h(t) = \lambda e^{\gamma t}$$

En la práctica, esta distribución es muy similar a la distribución Weibull pues también tiene un comportamiento monótono. En concreto, la función será creciente

Figura 3.4: **Distribución exponencial**

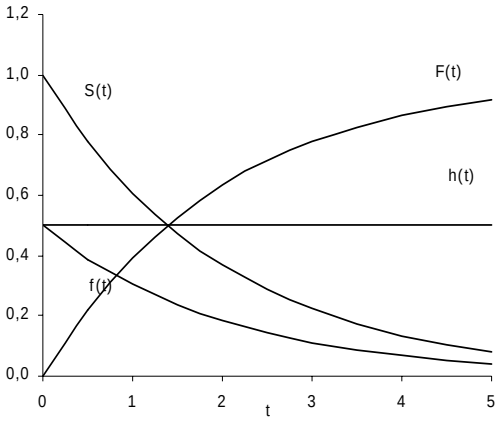


Figura 3.5: **Distribución Weibull**

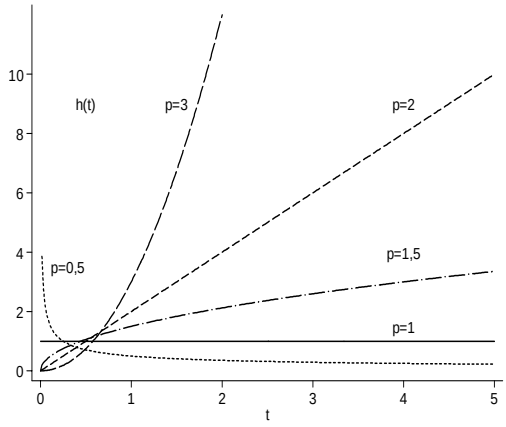


Figura 3.6: **Distribución Gompertz**

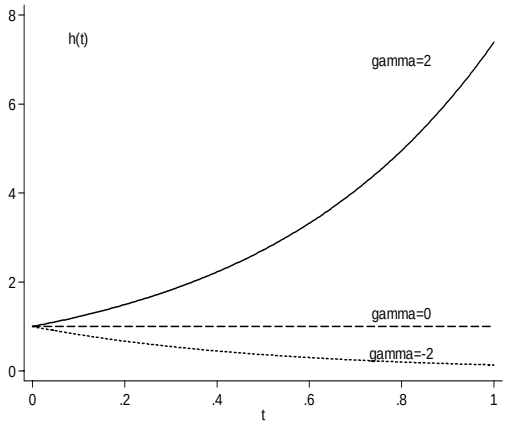


Figura 3.7: **Distribución Gamma**

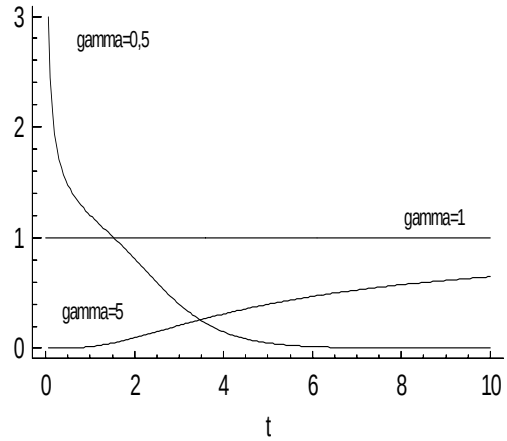


Figura 3.8: **Distribución log-normal**

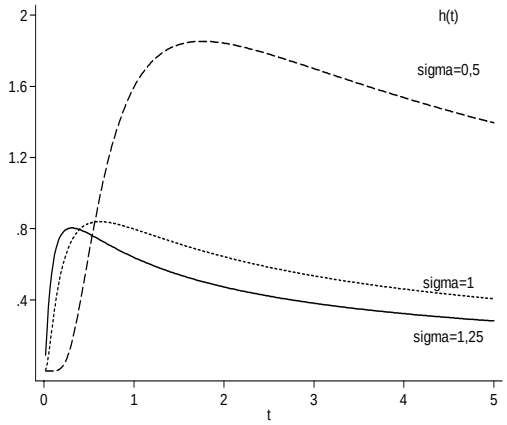
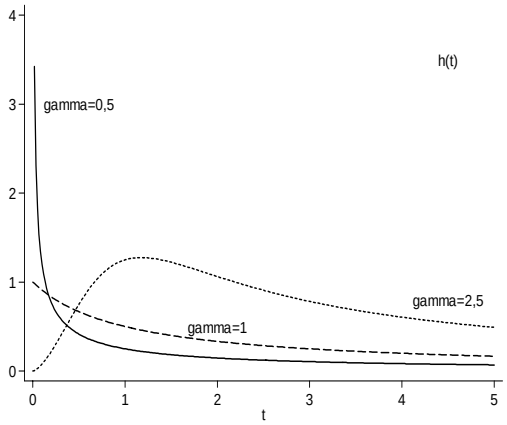


Figura 3.9: **Distribución log-logística**



cuando el parámetro γ sea superior a la unidad y decreciente en el caso contrario. Asimismo, la distribución exponencial es un caso particular de la Gompertz cuando $\gamma = 0$[24]. Su aportación a la modelización de duraciones se debe a que, frente a la distribución Weibull, el crecimiento (o decrecimiento) de la función de riesgo tiene caracter exponencial, lo que le ha permitido ser una firme candidata para describir la mortalidad de una población dentro del campo de la medicina.

El modelo de riesgos proporcionales en el que se incorpora el supuesto de que la función de riesgo base sigue una distribución Gompertz presenta la siguiente forma

$$h(t;x) = \lambda e^{\gamma t} e^{x\beta} = e^{\gamma t} e^{\beta_0 + x\beta}$$

donde una vez más uno de los parámetros de la distribución queda absorbido como término constante en la expresión que recoge el efecto de las variables explicativas.

Un buen ajuste a esta distribución supondría que los puntos correspondientes a las coordenadas $(\ln(H(t)), t)$ se aproximan a una recta.

Modelo logarítmico normal. Una variable sigue una distribución logarítmico normal si el logaritmo de esta variable sigue una distribución normal. En este caso, la distribución de T se distribuye como una logarítmico normal,

$$T \sim \text{Log-normal}(\beta_0, \sigma)$$

El modelo de duración logarítmico normal sólo puede ser interpretado como un modelo de riesgos acelerados,

$$\ln t = x\beta^* + \epsilon \tag{3.3}$$

donde la perturbación sigue una distribución normal de media 0 y varianza σ^2. La expresión $x\beta^*$ contiene un término constante para normalizar la perturbación de tal forma que su media sea igual a cero. Como se aprecia, la ecuación 3.3 es idéntica a la de un modelo de regresión clásico donde el término de error sigue una distribución normal, la única diferencia es que, en esta ocasión, algunas de las observaciones de la variable dependiente pueden estar censuradas y si estimáramos por mínimos cuadrados ordinarios este modelo, no se tendría en cuenta esta particularidad.

[24]En realidad, la distribución Gompertz está restringida a valores de γ superiores a cero dadas las implicaciones negativas que tiene el supuesto contrario, como señalan Klein y Moeschberger (1997).

La función de riesgo correspondiente a esta distribución tiene forma de joroba, esto es, se trata de una función unimodal y no monótona. En concreto empieza a crecer desde cero hasta alcanzar un máximo a partir del cual decrece. El principal inconveniente de esta distribución se deriva de que la función de riesgo decrece para valores grandes de t lo cual parece bastante improbable en muchos casos, aunque puede utilizarse si estos valores grandes no son realmente de interés en el estudio.

Modelo logarítmico logístico. La distribución logarítmico logística es muy similar a la distribución logarítmico normal, aunque presenta la ventaja de que la expresión de su función de supervivencia es más manejable. Asimismo, la expresión de la función de riesgo no es tampoco complicada,

$$h(t) = \frac{\lambda\gamma t^{\gamma-1}}{1+\lambda t^{\gamma}}$$

El comportamiento de esta función se deduce fácilmente una vez que se observa que el numerador coincide con la expresión de la función de riesgo de la distribución Weibull. Por lo tanto, su forma va a depender del valor del parámetro γ. En concreto, la función será monotóna decreciente si $0 < \gamma < 1$, o presentará forma de U invertida si $\gamma > 1$.

Modelo gamma generalizada. El modelo de duración gamma sólo puede expresarse como un modelo de riesgos acelerados en el que T sigue una distribución gamma de parámetros $(\lambda, \kappa, \sigma)$. Por lo tanto, la expresión de este modelo como un modelo de regresión viene dada por,

$$\ln t = x\beta^* + \epsilon$$

donde la perturbación sigue una distribución logarítmico gamma. La ventaja más importante de este modelo es su gran versatilidad que se deriva del hecho de contar con tres parámetros que permiten múltiples combinaciones para conseguir formas muy variadas en el comportamiento de la función de riesgo. Sin embargo, tiene el inconveniente de que la expresión de la función de riesgo es algo compleja lo que asusta en un principio a los investigadores sociales sin una fuerte base estadística. En contrapartida, su función de densidad es algo más tratable[25],

$$f(t) = \frac{\lambda\sigma^{\kappa}t^{\lambda\kappa-1}e^{-\sigma t^{\lambda}}}{\Gamma(\kappa)}$$

[25] En esta expresión, $\Gamma(\kappa)$ es la función gamma.

Además, el modelo gamma engloba algunos de los modelos ya expuestos sin más que imponer determinadas restricciones sobre los parámetros de la distribución. En concreto, la distribución gamma incluye como casos particulares la distribución Weibull ($\kappa = 1$), la exponencial ($\kappa = \sigma = 1$) y la logarítmico normal ($\kappa = 0$).

Tabla 3.4: **Función de densidad, supervivencia y riesgo de las distribuciones continuas habitualmente utilizadas en la modelización paramétrica**

DISTRIBUCIÓN	f. de riesgo $h(t)$	f. de supervivencia $S(t)$	f. de densidad $f(t)$	nº de parámetros
Exponencial	λ	$e^{-\lambda t}$	$\lambda e^{-\lambda t}$	1
Weibull	$\lambda p t^{p-1}$	$e^{-\lambda t^p}$	$\lambda p t^{p-1} e^{-\lambda t^p}$	2
Gompertz	$\lambda e^{\gamma t}$	$e^{[\frac{\lambda}{\gamma}(1-e^{\gamma t})]}$	$\lambda e^{\gamma t} e^{[\frac{\lambda}{\gamma}(1-e^{\gamma t})]}$	2
Log-normal	$\frac{f(t)}{S(t)}$	$1-\Phi(\frac{x-\mu}{\sigma})$	$1-\Phi(\frac{x-\mu}{\sigma})$	2
Log-logística	$\frac{\lambda\gamma t^{\gamma-1}}{1+\lambda t^{\gamma}}$	$\frac{1}{1+\lambda t^{\gamma}}$	$\frac{\lambda\gamma t^{\gamma-1}}{(1+\lambda t^{\gamma})^2}$	2
Gamma generalizada	$\frac{f(t)}{S(t)}$	$1-I(\sigma t^p, \kappa)$	$\frac{\lambda\sigma^{\kappa} t^{\lambda\kappa-1} e^{-\sigma t^{\lambda}}}{\Gamma(k)}$	3

De todos modos, en bastantes ocasiones es suficiente una versión reducida de esta distribución que depende únicamente de dos parámetros, la distribución *gamma estándar* que se obtiene haciendo el parámetro λ igual a cero. La función de densidad correspondiente a esta distribución viene dada por

$$f(t) = \frac{\sigma^{\kappa} t^{\kappa-1} e^{-\sigma t}}{\Gamma(\kappa)}$$

siendo σ el parámetro de escala y κ el parámetro de forma ($\sigma > 0$ y $\kappa > 0$). Entre sus propiedades cabe destacar: por un lado, se trata de una distribución con características muy similares a la distribución Weibull y por otro lado, se aproxima a la distribución normal cuando el parámetro κ tiende a infinito. Por su parte, la función de riesgo es monótona creciente para $\kappa > 1$ y monótona decreciente cuando $\kappa < 1$.

Para finalizar con esta exposición de los modelos paramétricos más conocidos, hemos recogido en la tabla 3.4, a modo de resumen, las expresiones de las funciones que caracterizan a estas distribuciones. En aquellos casos en los que la expresión resultaba demasiado compleja, se ha optado por no incluirla directamente en la tabla.

El modelo piece-wise constante. Se trata de un modelo que en cierta manera mimetiza al modelo semiparamétrico que veremos a continuación (el modelo de Cox), en el sentido de que es muy flexible en cuanto a la forma de la función de riesgo base. En concreto, si dividimos el tiempo en varios intervalos definidos como $(0, \tau_1]$, $(\tau_1, \tau_2]$, ..., $(\tau_1, \tau_2]$, $(\tau_{P-1}, \tau_P]$ y consideramos que la función se mantiene constante dentro de cada intervalo, la expresión del modelo *piece-wise* constante es

$$h(t;x) = \begin{cases} h_1(t)e^{x\beta} & t = (0, \tau_1] \\ h_2(t)e^{x\beta} & t = (\tau_1, \tau_2] \\ ... & \\ h_P(t)e^{x\beta} & t = (\tau_{P-1}, \tau_P] \end{cases}$$

donde $h_p(t) = \lambda_p \quad p = 1, 2, ..., P$.

Por lo tanto, en este caso se modeliza la función de riesgo base utilizando P parámetros, cada uno de ellos representando el riesgo base en un intervalo dado[26]. La estimación de este modelo se realiza por máxima verosimilitud y, frente a los modelos paramétricos vistos anteriormente, difiere en que se estima un mayor número de parámetros que vendrá determinado por el número de intervalos considerados. La principal dificultad en la implementación de esta alternativa se encuentra en la necesidad de fijar unos límites para cada intervalo, cuestión que no se basa en ningún criterio estadístico sino en la propia experiencia del investigador. Claramente, una elección juiciosa de los puntos de corte nos va a permitir realizar una aproximación razonable a cualquier función de riesgo base.

Estimación de los modelos paramétricos: máxima verosimilitud. El método de estimación de los modelos paramétricos consiste en la maximización de la función de verosimilitud. Ahora bien, en el proceso de construcción de la verosimilitud con datos de duración hemos de tener en cuenta que no todas las observaciones proporcionan la

[26] Una variante de este modelo es aquél en el que se supone que el logaritmo de la función de riesgo base se comporta linealmente dentro de cada intervalo, esto es, $\ln h_p(t) = \lambda_p + \gamma_p t$.

misma información debido a la existencia de censura en este tipo de datos. Por esta razón, hemos incluido este apartado en el que mostramos cuál es la contribución a la función de verosimilitud de las observaciones que están censuradas y, en particular, centraremos nuestra exposición en cómo se construye la función de verosimilitud en el caso de censura por la derecha[27]. La contribución a la función de verosimilitud de otros tipos de censura se recoge en la tabla 3.5 y, para un conocimiento más profundo de la estimación de estos modelos bajo otras clases de censura recomendamos la lectura del capítulo relativo a este tema en el manual de Klein y Moeschberger (1997).

Dadas las observaciones correspondientes a N individuos independientes $(t_i,\ \delta_i)$, donde t_i recoge la duración observada y δ_i es una variable dicotómica que informa si la observación está censurada ($\delta_i = 0$) o no ($\delta_i = 1$), la contribución a la función de verosimilitud de una observación no censurada vendrá recogida a través de su función de densidad $f(t_i)$. Por otro lado, en el caso de que la observación esté censurada por la derecha, la información que contiene dicha observación es que, al menos, la duración es igual a t_i. Consecuentemente, su contribución a la verosimilitud vendrá dado por el valor de la función de supervivencia en t_i. Por lo tanto, la función de verosimilitud de la muestra formada por los N individuos se puede escribir como,

$$L(\theta) = \prod_{i=1}^{N} (f(t_i, \theta))^{\delta_i} (S(t_i, \theta))^{1-\delta_i}$$

donde hemos denotado por θ el vector que contiene los parámetros a estimar. Aplicando neperianos sobre esta función,

$$\ln L(\theta) \quad = \quad \sum_{i=1}^{N} \delta_i \ln(f(t_i, \theta)) + \sum_{i=1}^{N} (1 - \delta_i) \ln(S(t_i, \theta))$$

y haciendo uso de las relaciones $f(t) = h(t)S(t)$ y $\ln S(t) = -H(t)$, se obtiene una expresión alternativa del logaritmo de la función de verosimilitud que enfatiza el porqué del uso de la función de riesgo dentro del análisis de duración,

$$\ln L(\theta) = \sum_{i=1}^{N} \delta_i \ln(h(t_i, \theta)) + \sum_{i=1}^{N} H(t_i, \theta),$$

[27]Damos por supuesto que la censura es no informativa lo que significa que no se requiere conocer la distribución de probabilidad de la censura para poder estimar por máxima verosimilitud los modelos paramétricos.

Una vez construida la función de verosimilitud se determina que valor de θ maximiza esta función siguiendo los pasos habituales de este procedimiento de estimación. El estimador de θ asi obtenido goza de las propiedades típicas de los estimadores máximo verosimiles, por lo tanto, será consistente y asintóticamente normal.

Tabla 3.5: **Contribución a la función de verosimilitud de distintas observaciones según el tipo de censura**

Tipo de observación	Contribución
duración exacta	$f(t)$
duración censurada por la derecha	$S(t)$
duración censurada por la izquierda	$1 - S(t)$
duración censurada en un intervalo	$S(t) - S(t-1)$
duración truncada por la izquierda	$f(t)/S(t_0)$

Nota: t es el tiempo censurado y t_0 el valor en que se trunca.

Fuente: Klein y Moeschberger (1997, pág. 66)

Validación. En ocasiones, es la misma teoría económica la que facilita la elección de uno o varios de los modelos paramétricos ya que se conoce cuál es el comportamiento de la función de riesgo (creciente, constante o decreciente). En el caso de un desconocimiento absoluto sobre la realidad que estamos intentando plasmar mediante un modelo paramétrico, entonces la decisión deberá ser tomada siguiendo criterios puramente estadísticos.

En concreto, podemos distinguir dos situaciones. Si los modelos están anidados, la utilización del test de la razón de verosimilitud o el test de Wald nos permitirá discriminar entre dos o varios modelos. En la breve descripción que hemos hecho de estos modelos se ha hecho referencia a esta cuestión y, en particular, el modelo más general de entre aquellos que aparecen en la tabla 3.4 es el que considera que T sigue una distribución gamma, derivándose a partir de éste, el modelo Weibull, el exponencial y el logarítmico normal.

Cuando los modelos no están anidados, es necesario recurrir a algún método de selección como es el criterio de información de Akaike (1975) o AIC. Este criterio propone penalizar el logaritmo de la verosimilitud con el número de parámetros que se están estimando en cada modelo. Concretamente, el estadístico AIC en este tipo de modelos

viene dado por

$$AIC = -2\log \text{ verosimilitud} + 2(c + p + 1)$$

donde c es el número de variables y p es el número de parámetros auxiliares del modelo. Dado que el mejor modelo es el que presenta un valor más alto en su función de verosimilitud, el modelo finalmente seleccionado será aquél que presente el valor más pequeño en el criterio comentado.

Otro procedimiento para seleccionar entre distintos modelos paramétricos hace uso de los residuos que se generan tras la estimación y, en particular, analiza los llamados residuos generalizados de *Cox-Snell*. Teniendo en cuenta que este procedimiento será también utilizado para evaluar la bondad del ajuste en la estimación semiparamétrica, no vamos a ser muy extensos en este apartado. La técnica consiste en obtener una estimación empírica de la función de riesgo acumulada, por ejemplo la proporcionada por el estimador de Kaplan-Meier, considerando que los residuos Cox-Snell hacen el papel de la variable duración pero manteniendo la misma variable indicadora de la censura. El paso siguiente es representar gráficamente esta función de riesgo acumulada frente a los residuos y analizar en qué forma está dispuesta la nube de puntos. En el caso de un buen ajuste, los puntos deberían encontrarse próximos a la bisectriz en cuyo caso se podría decir que el modelo paramétrico se aproxima bastante a los datos.

3.2.2.3 Modelización semiparamétrica: el modelo de Cox

La modelización paramétrica es preferible cuando la distribución elegida es la correcta. Sin embargo, no es muy normal que el investigador económico pueda sustentar dicha elección basándose en los postulados de la teoría económica. En este contexto, la modelización semiparamétrica propuesta por Cox (Cox, 1972), que vamos a explicar a continuación, puede ser una opción de gran utilidad.

En realidad, el modelo de Cox, también conocido como modelo de *regresión* de Cox, no es más que un modelo de riesgos proporcionales. Por lo tanto, las variables explicativas tienen un efecto multiplicativo sobre la función de riesgo base, de tal forma que la función de riesgo se construye como el producto de una función de riesgo base, $h_0(t)$, por una función que depende tan sólo de las variables explicativas, $\varphi(x)$,

$$h(t; x) = h_0(t)\varphi(x)$$

Suponiendo que $\varphi(x) = e^{x\beta}$, el modelo de Cox presenta la siguiente estructura,

$$h(t;x) = h_0(t)e^{x\beta}$$

Estimación del modelo de Cox: máxima verosimilitud parcial El artículo seminal de Cox (1972) supuso un hito en el análisis de supervivencia pues a la par que planteó un modelo que en la actualidad es el más utilizado, también propuso un ingenioso método de estimación que se conoce hoy en día como estimación por máxima verosimilitud parcial.

De cara a estudiar las características de este método de estimación, vamos a suponer que disponemos de información sobre la duración de un suceso para N individuos, de las cuales $N-r$ son censuradas y no se repiten y, vamos a denotar con $t_{(1)} < t_{(2)} < ... < t_{(r)}$ a las r duraciones ordenadas.

El planteamiento de la estimación de estos modelos frente a la estimación por máxima verosimilitud difiere en que en lugar de construir la verosimilitud a partir de la pregunta sobre cuál es la probabilidad de que la duración correspondiente al individuo i sea t_i, se considera más bien cuál es la probabilidad de que la duración t_j corresponda al individuo i sabiendo que se ha observado a un individuo con duración t_j. Esta probabilidad condicionada se puede expresar como,

$$\frac{h(t_{(j)};\beta)}{\sum\limits_{i\in R(t_{(j)})} h(t_{(j)};\beta)}$$

si la función h cumple la hipótesis de riesgos proporcionales se obtiene

$$\frac{h_0(t_{(j)})e^{x_{(j)}\beta}}{h_0(t_{(j)})\sum\limits_{i\in R(t_{(j)})} e^{x_{(i)}\beta}} = \frac{e^{x_{(j)}\beta}}{\sum\limits_{i\in R(t_{(j)})} e^{x_{(i)}\beta}}$$

expresión en la que el término $h_0(t)$ queda cancelado. En definitiva, la función de verosimilitud parcial presenta la siguiente forma

$$L_p(\beta) = \prod_{j=1}^{r} \frac{e^{x_{(j)}\beta}}{\sum\limits_{i\in R(t_{(j)})} e^{x_{(i)}\beta}}$$

que se caracteriza porque sólo los individuos cuyas duraciones no están censuradas contribuyen explícitamente a la función de verosimilitud aunque sí se tienen en cuenta las duraciones no censuradas de forma implícita en el cálculo del denominador $\sum\limits_{i\in R(t_{(j)})} e^{x_{(i)}\beta}$,

cuyo sumatorio incluye a todos los individuos que están sujetos a riesgo en $t_{(j)}$ y que hemos denotado por $R(t_{(j)})$[28].

La ventaja de este novedoso procedimiento es que permite obtener estimaciones de los coeficientes β sin necesidad de especificar la función de riesgo base $h_0(t)$. Este método de estimación ha tenido gran aceptación entre aquellos investigadores más preocupados por determinar el efecto de las variables explicativas sobre la duración del suceso que por la forma de la función de riesgo.

Los estimadores obtenidos mantienen la mayor parte de las propiedades que cumplen los estimadores máximo verosímiles. En concreto, son consistentes y asintóticamente normales aunque no serán los más eficientes.

Una vez estimado el modelo de Cox, es posible obtener estimadores de la función de supervivencia y de la función de riesgo acumulada recurriendo a las siguientes expresiones,

$$\hat{S}(t;x) = [\hat{S}_0(t)]^{e^{x\hat{\beta}}}$$

$$\hat{H}(t;x) = \hat{H}_0(t)e^{x\hat{\beta}}$$

si bien para su cálculo es necesario obtener previamente estimadores de las funciones base ($S_0(t)$ y $H_0(t)$). Los principios que rigen el cálculo de estas funciones base están muy relacionados con los que estudiamos al construir el estimador de Kaplan-Meier pero aquí no nos vamos a detener en ello (véase Singer y Willet, 2003). Sí haremos hincapié en un hecho de gran trascendencia en las aplicaciones del modelo de Cox, y es que no se estima directamente la función de riesgo base, aunque se podría obtener una estimación de la misma mediante la utilización de algún método de suavizado (por ejemplo, el método kernel) sobre la función de riesgo acumulada.

Tratamiento de los empates. Una dificultad añadida al estimar el modelo de Cox es el tratamiento de los "empates". En nuestra exposición sobre la forma en que se estima el modelo de Cox se ha explicado con cierto detalle en qué consiste el método de verosimilitud parcial, cuya principal ventaja radica en la posibilidad de estimar los efectos de las variables explicativas sin necesidad de especificar la forma funcional de la función de riesgo base. Esto se debe principalmente a que en este procedimiento

[28] El término *parcial* alude, en este caso, a la utilización de tan sólo parte de la información disponible, en concreto no se tiene en cuenta el valor de la duración directamente sino tan sólo su orden o rango.

estadístico de estimación sólo se tiene en cuenta el orden de los valores de la duración y no los valores en sí. Esta ventaja se puede tornar en un problema cuando los valores de T se repiten de forma reiterada, dado que impide que se les asigne un orden determinado. En la literatura se pueden encontrar diferentes soluciones a este problema, algunas de ellas son tan sólo aproximaciones mientras que otras, aunque requieren de una ingente cantidad de cálculos, nos conducen a estimaciones exactas.

El primero de los métodos exactos, denominado *método marginal*, considera que la existencia de empates se debe a la falta de precisión al observar la variable, y por lo tanto, que existe un orden desconocido entre las observaciones empatadas. De esta manera, la contribución a la verosimilitud de cada observación $t_{(j)}$ empatada m veces se obtiene evaluando su valor para cada una de las m! posibles ordenaciones que podrían darse. Por ejemplo, supongamos que tenemos dos individuos que han estado dos meses desempleados. En realidad, posiblemente ninguno de los dos haya estado exactamente ese tiempo desempleado pero no hemos sido capaces de obtener una información más precisa. En este caso, si calculamos la contribución de ese valor de duración, $t_{(j)} = 2$, a la función de verosimilitud tenemos que considerar todas las posibles situaciones que pueden haber generado esas dos duraciones: que haya salido primero el individuo 1 y luego el individuo 2, o la situación contraria.

El otro método exacto, propuesto por el mismo Cox (1972) en su artículo original, no ha tenido gran aceptación principalmente porque los programas informáticos no suelen implementarlo. En algunos ámbitos, ha recibido el nombre de *método parcial* y parte del principio de que los empates no son el resultado de una imprecisión en la medición del tiempo sino que el proceso en sí es realmente discreto.

El gran inconveniente de estos dos métodos exactos es el enorme consumo de tiempo que supone su utilización, de ahí que, se hayan popularizado otros métodos aproximados para obtener la función de verosimilitud en el caso de empates. Dentro de los métodos aproximativos se encuentran el método de Breslow (1974) propuesto inicialmente por Peto (1972) -que es el que viene implementado por defecto en la mayoría de los programas estadísticos- y el método de Efron (1977). Los dos métodos reducen el número de cálculos que supone la aplicación del método marginal, realizando una corrección en el denominador de la expresión de la contribución a la verosimilitud de cada una de las observaciones empatadas. En concreto en el método de Breslow se mantiene el mismo denominador para todos los valores empatados, incluyendo todos los términos cuya duración es igual o superior a $t_{(j)}$, mientras que en el método de Efron se utiliza

una ponderación.

La aproximación de Breslow funciona bastante bien cuando el número de empates es reducido, y en concreto, cuando el número de empates en relación al conjunto de individuos sujetos a riesgo es pequeño para cada $t_{(j)}$. La utilización de este método en estos casos conduce a coeficientes muy similares a los que se obtendrían con el método exacto y los tests de restricciones sobre los parámetros nos llevarían a conclusiones semejantes. En el caso contrario, esto es, cuando los empates se concentran en determinados valores, los coeficientes estarán sesgados hacia cero. Por lo tanto, cuando el número de empates es importante, se recomienda utilizar el método de Efron, siempre que el método exacto no sea viable (Allison, 1995).

Validación. En cualquier modelo econométrico tras su estimación es necesario someter al mismo a diversas pruebas estadísticas que permitan comprobar su validez y calidad. En este sentido, en el caso del modelo de Cox existen en la literatura diferentes pruebas de diagnosis relativas, por un lado, a la especificación de la parte $e^{x\beta}$ (selección y forma funcional de las variables explicativas) y, por otro lado, a contrastar las hipótesis que subyacen en el modelo, como es la hipótesis de proporcionalidad.

Es práctica habitual en la validación de un modelo obtener una medida del error cometido, lo que en los modelos de regresión se denomina residuo. La construcción de unos residuos en el modelo de Cox, como veremos, no tiene una única solución ya que se trata de un modelo con unas características muy específicas a las que hemos ido aludiendo a lo largo de este capítulo. Las razones que dificultan la definición de un residuo tal y como se entiende en los modelos de regresión, esto es, como diferencia entre el valor observado y el valor ajustado, son: la existencia de datos censurados, la propia especificación del modelo en términos de la función de riesgo y el método de estimación. En concreto, los residuos pierden su sentido habitual cuando los valores observados se encuentran censurados y los valores estimados no se corresponden con estimaciones de los valores esperados de la variable.

Salvando estos inconvenientes, encontramos en la literatura una amplia gama de "residuos" que evalúan diferentes aspectos del modelo estimado. Entre estas propuestas resaltamos, por un lado, los residuos Cox-Snell, los residuos martingala y los residuos deviance -todos ellos basados en la función de riesgo acumulada- y, por otro lado, los residuos Schoenfeld y los residuos score -construidos a partir de la función de verosimilitud, aunque aquí sólo trataremos de explicar el funcionamiento de los que

se utilizan más habitualmente.

Es bien sabido que dada una variable aleatoria X con función de supervivencia $S(x)$, la transformación $Y = -\ln S(X)$ tiene una distribución exponencial con media uno, independientemente de la forma de $S(x)$. En base a este resultado, se definen los *residuos Cox-Snell* como

$$e_i = -\ln(\widehat{S}(t_i; x)) = \widehat{H}(t_i; x)$$

y sustituyendo,

$$e_i = \widehat{H}_0(t_i)e^{x_i\widehat{\beta}}$$

Por lo tanto, los valores de estos residuos son realizaciones de una distribución exponencial de media uno si el modelo es el correcto. Siendo más precisos, y dado que trabajamos con datos censurados, son una muestra censurada de esta distribución. El hecho de conocer la distribución de estos residuos nos va a permitir valorar de forma global el modelo estimado. El procedimiento es el siguiente. Una vez obtenidos los residuos se calcula la función de riesgo acumulada -por ejemplo, utilizando el estimador no paramétrico de Nelson-Aalen- correspondiente a los residuos ($H(e_i)$) y trabajando con la misma variable de censura. A continuación se representan los valores de $H(e_i)$ frente a e_i. Si el modelo ajustado es el correcto, los puntos estarán situados próximos a una recta que pasa por el origen y con pendiente de 45º.

Otro procedimiento que nos permite evaluar la especificación general del modelo -no específico de estos modelos- es una variante del test RESET de Ramsey (1969) que consiste en estimar un nuevo modelo de Cox en el que se incluyen términos de potencias y productos cruzados de las variables explicativas o, lo que es lo mismo, potencias de la variable endógena estimada. En este modelo se contrasta la restricción de que los coeficientes que acompañan a estas potencias son iguales a cero.

Dentro del segundo bloque de residuos, hablamos de los *residuos Schoenfeld*, nombre que reciben debido a su precursor Schoenfeld (1982), y que se obtienen al sustituir el parámetro β_k por su estimador ($\widehat{\beta}_k$) en la expresión de la derivada parcial de la función del logaritmo de la verosimilitud parcial con respecto a dicho parámetro. Por lo tanto, siendo la expresión de esa derivada parcial,

$$\frac{\partial \ln L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} \delta_i \left(x_{ik} - \frac{\sum_{j \in R(t_i)} x_{jk} e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right)$$

el residuo Schoenfeld correspondiente al individuo i y al regresor k se define como

$$s_{ik} = \delta_i \left(x_{ik} - \frac{\sum_{j \in R(t_i)} x_{jk} e^{x_j \widehat{\beta}}}{\sum_{j \in R(t_i)} e^{x_j \widehat{\beta}}} \right) \qquad i = 1, ..., N \quad k = 1, ..., K$$

donde δ_i es igual a 0 si la observación está censurada. La expresión anterior compara el valor de la variable X en el momento en que acontece el suceso para el individuo i (t_i) con una media ponderada de los valores de dicha variable correspondientes a los individuos que están sujetos a riesgo en t_i, siendo la ponderación $e^{x_j \widehat{\beta}}$.

Los residuos Schoenfeld se utilizan para validar la hipótesis de proporcionalidad. Como mencionamos en la introducción de este apartado, en el modelo de Cox hay implícitas una serie de hipótesis siendo la de proporcionalidad de los riesgos una de las fundamentales, y sin la cual el modelo perdería todas sus ventajas frente a otros modelos de duración. Por ello, el análisis del cumplimiento de esta hipótesis resulta imprescindible en cualquier investigación.

Debido a que los residuos Schoenfeld no dependen de t, éstos se convierten en la herramienta básica para llevar a cabo esta labor. De hecho, si en una regresión de estos residuos frente al tiempo se comprueba que el parámetro es distinto de cero, la sospecha de que la variable en cuestión no verifica la hipótesis de proporcionalidad se convierte en certeza. En la práctica, los investigadores prefieren la utilización de un procedimiento gráfico en el que aparecen los residuos Schoenfeld frente al tiempo para observar si éstos muestran un comportamiento determinado. Bajo la hipótesis nula de proporcionalidad, los residuos deben situarse entorno a una recta de pendiente igual a cero.

Dada la importancia del cumplimiento de la hipótesis de proporcionalidad, existen en la literatura numerosos procedimientos para determinar la validez de la misma. Básicamente, vamos a distinguir tres tipos de procedimientos: aquéllos basados en representaciones gráficas de las funciones de supervivencia, aquéllos que plantean la inclusión de nuevas variables que no son más que interacciones de las variables del modelo con el tiempo y por último, métodos basados en la medición de la bondad del ajuste a través del análisis de los residuos -ya comentados.

Con respecto al primer procedimiento, éste consiste en representar en una mismo gráfico la función $\ln(-\ln(\hat{S}(t;x)))$ para diferentes categorías de la variable X cuya proporcionalidad se está analizando, siendo $\hat{S}(t;x)$ una estimación de la función de supervivencia (por ejemplo, la obtenida por el método de Kaplan-Meier). Si las curvas son paralelas entre sí, la hipótesis de proporcionalidad sería aceptable.

Por otro lado, dado que la hipótesis de proporcionalidad implica que el efecto de cada variable es constante en el tiempo, la violación de dicha hipótesis es equivalente a considerar que la interacción de una (o varias) variable(s) y el tiempo es significativa. En base a esto, una forma muy sencilla de evaluar la hipótesis de proporcionalidad consiste en incluir este tipo de interacciones y contrastar su significatividad. El modelo, tras la inclusión de las interacciones, vendría especificado como

$$h(t;x) = h_0(t)e^{x\beta + xg(t)\gamma}$$

donde $g(t)$ puede ser cualquier función de t, lo que hace que la detección del incumplimiento de la hipótesis de proporcionalidad dependa de la función $g(t)$ elegida, dificultando la puesta en práctica de este método de evaluación. En general, en las aplicaciones se suele utilizar t o $\ln t$.

Cuando alguna variable claramente no acepta la hipótesis de proporcionalidad, una posible solución es estimar un modelo de Cox *estratificado* por la variable en cuestión, el cual se caracteriza porque se establece una forma de la función de riesgo diferente para cada estrato, aunque los efectos de las variables siguen siendo los mismos para todos los estratos, esto es

$$h_s(t;x) = h_{0s}e^{x\beta}$$

donde el subíndice s representa un estrato determinado. El principal inconveniente de este modelo es que impide la cuantificación del efecto sobre la duración de la variable por la cual se estratifica.

3.3 Análisis de duración en tiempo discreto

3.3.1 Introducción

En las investigaciones empíricas, los modelos de duración en tiempo continuo siguen prevaleciendo frente a los modelos en tiempo discreto. Sin embargo, el tratamiento discreto de los modelos de duración, como veremos, va a resultar de gran utilidad en determinadas situaciones. Varios autores se han acercado a este tema entre los que resaltamos por sus aportaciones: Prentice y Gloecker (1978), Allison (1982), Meyer (1990, 1995), Kiefer (1988, 1990), Singer y Willett (1993, 2003) y Jenkins (1995, 2004).

Como ya advertimos en la introducción de este capítulo, en la práctica nos encontramos con la necesidad de disponer de modelos de duración en tiempo discreto.

Las razones son varias. En primer lugar, la dificultad de medir el tiempo de forma precisa conduce a observaciones discretas de una variable continua. Por ejemplo, aquellos estudios que recurren a encuestas periódicas, construyen la duración a partir de la información de dos entrevistas consecutivas. Si entre dos entrevistas el individuo ha cambiado de estado, el valor asignado a la duración estará comprendido dentro de un intervalo cuya amplitud estará condicionada por la periodicidad de la encuesta. A este tipo de datos nos referiremos como *datos de duración agrupados* en intervalos. En segundo lugar, la presencia de un número de empates elevado en la variable duración, como ya hemos mencionado, supone que la estimación del modelo en tiempo continuo se convierta en una tarea realmente ardua. La existencia de empates puede deberse a la tendencia de los entrevistados a redondear en aquellas encuestas de carácter retrospectivo en las que el individuo tiene que recurrir a su memoria para contestar a la pregunta referida a la duración. Por último, la ocurrencia de determinados sucesos, por su propia naturaleza, está limitada a un tiempo discreto. Este sería el caso de una investigación sobre la duración de los estudios universitarios, en la que la ocurrencia del suceso sólo tiene lugar en fechas muy concretas. En este caso hablaríamos de *datos de duración discretos*.

De lo expuesto en el párrafo anterior se deduce que dentro de los modelos de duración discretos podemos encontrar dos enfoques. Por un lado, aquél que estipula que la duración es realmente una variable discreta. Y, de forma alternativa, aquél en el que se considera que la duración es continua pero ha sido observada de forma discreta, esto es, la información viene agrupada en intervalos. Dado que ambas alternativas conducen a resultados idénticos y tan sólo se diferencian en las formas, aquí vamos a explicar el caso en el que se trabaja con datos estrictamente discretos.

3.3.2 Función de supervivencia y función de riesgo

El supuesto de que la duración es una variable discreta nos obliga a redefinir las principales funciones que la caracterizan.

Sean t_1, t_2, ..., t_j, ..., t_J, los valores discretos ordenados que puede tomar la variable T. Definimos entonces la función de supervivencia como la probabilidad de que el suceso no haya aún acontecido en t_j, esto es,

$$S(t_j) = p(T > t_j)$$

Claramente el concepto que aparece recogido en esta expresión no se ha visto modificado por la consideración de que la variable es discreta y la única diferencia con respecto al caso continuo radica en la propia variable T y, en concreto, en sus posibles valores. En consecuencia, las propiedades que señalabamos con respecto a esta función en los modelos de duración en tiempo continuo se van a mantener, y, por lo tanto, la función será monótona decreciente y tomará valores comprendidos dentro del intervalo $[0, 1]$.

Respecto de la función de riesgo, esta función valora la probabilidad de que ocurra el suceso en el periodo t_j condicionada a que el suceso no ha tenido lugar hasta entonces, probabilidad que queda recogida matemáticamente en la siguiente expresión:

$$h(t_j) = p(T = t_j \,/\, T \geq t_j)$$

Observamos que la formulación matemática de la función de riesgo sí que se ha visto modificada. Además esta modificación tiene implicaciones más importantes que las meramente de definición, pues al contrario que en el campo continuo, ahora sí se trata de una probabilidad y como tal varía entre 0 y 1.

Por último, la función de densidad, que informa sobre la probabilidad de ocurrencia del suceso en t_j, tendrá la siguiente expresión,

$$f(t_j) = p(T = t_j)$$

A continuación presentamos ciertos resultados que relacionan estas funciones en el campo discreto. Estos resultados nos serán de gran utilidad a la hora de plantear la estimación de estos modelos. En concreto, la probabilidad de que la duración sea superior a t_j se puede descomponer en el producto de una serie de términos, uno por periodo, que recogen las probabilidades condicionadas de no ocurrencia del suceso en ninguno de los periodos anteriores a t_j, esto es, la función de supervivencia se puede expresar como:

$$\begin{aligned} S(t_j) &= p(T > t_j) \\ &= p(T \neq t_j / T \geq t_j) \cdot (T \neq t_{j-1} / T \geq t_{j-1}) \cdot \ldots \cdot p(T \neq t_1 / T \geq t_1) \\ &= (1 - h(t_j)) \cdot (1 - h(t_{j-1})) \cdot \ldots \cdot (1 - h(t_1)) \\ &= \prod_{s=1}^{j} (1 - h(t_s)) \end{aligned}$$

De igual modo, la probabilidad de que la duración sea igual a t_j se puede descomponer en el producto de las probabilidades condicionadas de que no ocurra del suceso en ninguno de los periodos intermedios y sí en el último periodo, lo que nos lleva a expresar la función de densidad como:

$$\begin{aligned} f(t_j) &= p(T = t_j) \\ &= p(T = t_j/T \geq t_j) \cdot p(T \neq t_{j-1}/T \geq t_{j-1}) \cdot ... \cdot p(T \neq t_1/T \geq t_1) \\ &= h(t_j) \cdot (1 - h(t_{j-1})) \cdot ... \cdot (1 - h(t_1)) \\ &= h(t_j) \prod_{s=1}^{j-1} (1 - h(t_s)) \end{aligned}$$

3.3.3 Especificación y estimación

La estimación de estos modelos se realiza maximimizando la función de verosimilitud. Como de costumbre, en los datos de duración nos podemos encontrar con dos tipos de contribución a la verosimilitud: a) la de aquellas observaciones que presentan censura por la derecha ($\delta_i = 0$) para las cuales la información que tenemos es que la duración es superior a t_j -función de supervivencia en t_j- y b) la correspondiente a las observaciones no censuradas ($\delta_i = 1$) cuya contribución vendrá determinada por la probabilidad de que el suceso ocurra en t_j -función de densidad en t_j. En definitiva, la función de verosimilitud correspondiente a una población formada por N individuos homogéneos tendrá la siguiente forma,

$$L = \prod_{i=1}^{N} p(T_i = t_j)^{\delta_i} p(T_i > t_j)^{(1-\delta_i)}$$

y sustituyendo los resultados antes obtenidos para la función de densidad y para la función de supervivencia,

$$L(\theta) = \prod_{i=1}^{N} \left[h(t_{ij}, \theta) \prod_{s=1}^{j-1} (1 - h(t_{is}, \theta)) \right]^{\delta_i} \left[\prod_{s=1}^{j} (1 - h(t_{is}, \theta)) \right]^{(1-\delta_i)} \tag{3.4}$$

Para poder maximizar esta expresión se requiere especificar previamente el comportamiento de la función de riesgo, $h(t_j, \theta)$. En la literatura se han considerado dos especificaciones. La primera, que nos conduce al *modelo log-log del complementario*, es una representación discreta del modelo de riesgos proporcionales en tiempo continuo -como asi demuestran Prentice y Gloecker (1978)- y, por lo tanto se adapta mejor a situaciones en las que la duración es una variable continua observada de forma

discreta, esto es, en el caso de datos de duración agrupados en intervalos. La segunda especificación, que se conoce por el *modelo logístico*, fue planteada inicialmente por Cox (1972). Este autor propuso esta especificación basándose en que la función de riesgo del modelo discreto recoge probabilidades, por lo que es lógico parametrizar esta función, de forma análoga a los modelos de elección discreta, mediante una función logística. Este modelo no es un modelo de riesgos proporcionales.

La función de riesgo en el *modelo log-log del complementario*, dadas unas características del individuo x, viene definida por

$$h(t_j; x) = 1 - e^{-e^{\theta_j + x\beta}}$$

expresión que se deriva tras plantear la estimación de un modelo continuo con datos agrupados en intervalos[29]. El nombre por el que se conoce a este modelo se debe a la función que se obtiene al formularlo como un modelo lineal, como se deduce de la siguiente expresión

$$\ln(-\ln(1 - h(t_j; x))) = \theta_j + x\beta \tag{3.5}$$

Por otro lado, la utilización de la función logística para parametrizar la función de riesgo nos lleva a la siguiente expresión:

$$\ln\left[\frac{h(t_j; x)}{1 - h(t_j; x)}\right] = \theta_j + x\beta \tag{3.6}$$

que nos permite ver claramente el incumplimiento de la proporcionalidad de los riesgos ya que en esta ecuación se está asumiendo que las variables explicativas están linealmente asociadas con la transformación logística de la función de riesgo, no con la función de riesgo directamente ni con el logaritmo de esta función. Se deriva sin dificultad la función de riesgo de este modelo

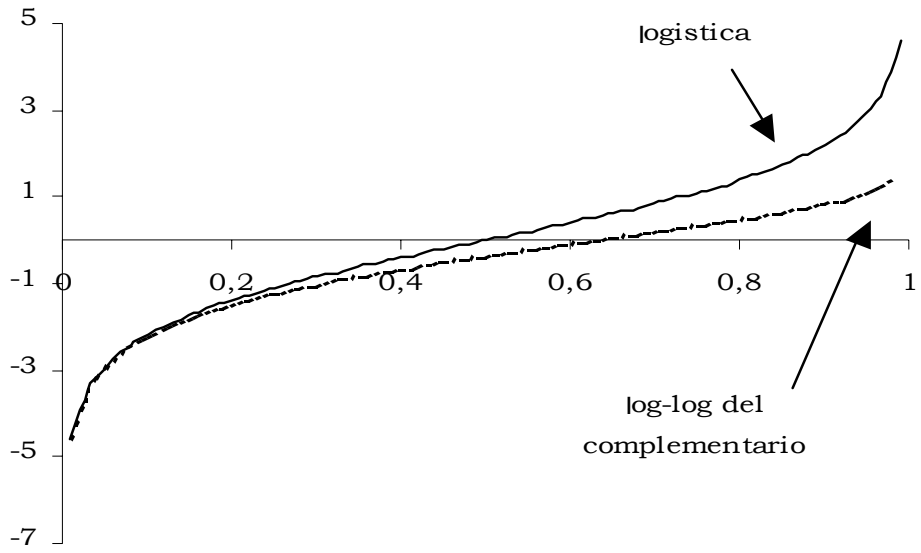
$$h(t_j; x) = \frac{1}{1 + e^{-\theta_j - x\beta}}$$

En la práctica, la dos especificaciones dan lugar a resultados muy similares ya que el modelo logístico converge al modelo log-log del complementario cuando los valores de la función de riesgo son muy pequeños, tal y como se advierte al observar los gráficos de

[29] Esta no es la única especificación que es consistente con un modelo de duración en tiempo continuo (Sueyoshi, 1995) aunque sí la más utilizada.

ambas distribuciones que aparecen en la figura 3.10. En dicha figura podemos apreciar que ambas funciones restringen los valores de $h(t)$ -tal y como era de esperar- a valores comprendidos entre cero y uno. Ahora bien, mientras la función logística es simétrica, la función log-log del complementario presenta cierta asimetría. Por lo tanto, la elección de una de las especificaciones vendrá más bien determinada por cuestiones referidas al cumplimiento de la hipótesis de proporcionalidad.

Figura 3.10: **Representación gráfica de la función log-log del complementario y de la función logística**



Las dos especificaciones contienen dos tipos de parámetros: los θ's que captan el nivel de riesgo base en cada periodo de tiempo y los β's que recogen el efecto de cada x sobre la función de riesgo. Respecto de los primeros, el valor y la dirección de su variación nos describen la forma de la función de riesgo e informan de si el riesgo crece, decrece o se mantiene constante en el tiempo. En principio, tal y como muestran las ecuaciones 3.5 y 3.6 no existe un único término constante en estos modelos sino que aparecen tantos términos como número de periodos se haya considerado al dividir el tiempo. Claramente, para recoger esta característica del modelo tan sólo es necesario construir una serie de variables dicotómicas para cada valor de t. Atendiendo a esta

consideración, el modelo logístico presentaría la siguiente expresión,

$$\ln\left[\frac{h(t_j;x)}{1-h(t_j;x)}\right] = \theta_1 D_1 + \theta_2 D_2 + ... + \theta_J D_J + x\beta$$

En estas circunstancias, el modelo se caracteriza porque no impone ninguna restricción al comportamiento de la función de riesgo, lo que en cierta medida le hace asemejarse al modelo semiparamétrico o modelo de Cox anteriormente analizado. La ventaja de los modelos discretos es que nos permiten obtener una estimación de la función de riesgo base directamente.

Otra posibilidad es asumir un determinado comportamiento de la función de riesgo base tal y como hacemos en los modelos paramétricos (exponencial, Weibull, etc.), para ello basta con introducir una determinada función de t, en lugar de las variables dicotómicas, en cualquiera de las especificaciones[30].

Una vez que la función de riesgo ha quedado especificada, la estimación por máxima verosimilitud nos permitiría obtener estimadores de los parámetros β y de la función de riesgo base. Sin embargo, tal como demuestra Allison (1982), la estimación de estos modelos es mucho más sencilla si reescribimos la función de verosimilitud incorporando en la expresión una variable dicotómica definida para cada periodo j como,

$$Y_{ij} = \begin{cases} 1 & t_{is} = t_{ij} \\ 0 & t_{is} < t_{ij} \end{cases}$$

Por ejemplo, en el caso de un individuo que está desempleado tres meses tendríamos tres valores, uno por cada periodo, siendo los dos primeros iguales a cero y el correspondiente al último mes -en el que se produce la salida del desempleo- igual a uno. Si la duración está censurada, por ejemplo cuando al individuo se le pierde tras haber permanecido ocho meses desempleado, el valor será igual a cero en cada uno de los ocho meses en que se divide el tiempo. Al obtener la expresión de la función de verosimilitud en términos de esta nueva variable,

$$L(\theta) = \prod_{i=1}^{N} \prod_{s=1}^{j} h(t_{ij};\theta)^{Y_{is}} (1 - h(t_{ij};\theta))^{1-Y_{is}}$$

se observa que tiene la misma forma que la función de verosimilitud de un modelo de elección discreta binario en el que la variable endógena es Y_{ij}, lo que facilita

[30] La distinción entre modelos paramétricos y semiparamétricos que consideramos en los modelos de duración continuos no es tan relevante en este caso ya que en una misma especificación -logística o log-log del complementario- se pueden incluir las dos situaciones.

considerablemente la obtención de los estimadores máximo verosimiles mediante la utilización de cualquier programa estadístico en el que aparezca implementado este tipo de modelos de elección discreta. Si se opta por construir esta función de verosimilitud en lugar de la que aparece en la ecuación 3.4, es necesario hacer una reorganización de los datos de tal manera que se pase de tener una obervación por individuo a tener tantas observaciones por individuo como veces éste haya estado sujeto a riesgo. En la reorganización de los datos se han de construir dos nuevas variables, una que identifique al individuo y otra que distinga el periodo que se está analizando dentro del conjunto de periodos en los cuales el individuo ha estado sujeto a riesgo.

Tabla 3.6: **Muestra de la estructura de los datos antes y después de la expansión necesaria para poder estimar un modelo discreto**

Estructura inicial de los datos

id	duración	censura	edad	sexo
1	3	1	34	0
2	8	0	19	1

Estructura después de haber expandido los datos

id	duración	censura	edad	sexo	T	Y
1	3	1	34	0	1	0
1	3	1	34	0	2	0
1	3	1	34	0	3	1
2	8	0	19	1	1	0
2	8	0	19	1	2	0
2	8	0	19	1	3	0
2	8	0	19	1	4	0
2	8	0	19	1	5	0
2	8	0	19	1	6	0
2	8	0	19	1	7	0
2	8	0	19	1	8	0

En la tabla 3.6 se presenta un ejemplo de como quedarían los datos tras expandir la muestra. En la primera parte de esta tabla aparecen los datos iniciales correspondientes

a dos individuos. El primer individuo, un varón de 34 años ha estado desempleado 3 meses. La segunda duración, censurada por la derecha, corresponde a una mujer de 19 años que ha estado desempleada durante al menos 8 meses pues tras ese tiempo se la ha dejado de observar. En la segunda parte de la tabla vemos como se transforman los datos, obteniendo esa variable binaria Y y una nueva variable que indica el periodo en el que se encuentra el individuo. De esta manera, para el primer individuo, tras la reorganización de los datos, tenemos tres observaciones, tantas como veces ha estado sujeto al riesgo de encontrar ese primer empleo significativo. Mientras que para la joven el número de observaciones es igual a ocho, una por cada mes en el que ha sido observada y ha estado desempleada.

3.4 Variables dependientes del tiempo

3.4.1 Introducción

Hasta este punto se ha considerado por razones de simplicidad que los valores de las variables explicativas incorporadas en los modelos analizados no cambiaban durante todo el periodo de observación. Sin embargo, el análisis dinámico que se realiza a través de las técnicas englobadas en los modelos de duración posibilita el tratamiento de este tipo de variables, lo que mejora sustancialmente el análisis.

En esta sección vamos a explicar como incluir estas variables en los modelos y cuáles son los principales problemas asociados a este tema. Retomamos por su gran repercusión en las aplicaciones prácticas el modelo de Cox aunque la incorporación de variables dependientes del tiempo en cualquiera de los modelos expuestos hasta ahora presenta, en su gran mayoría, complicaciones no muy diferentes de las que comentaremos para el modelo de Cox.

3.4.2 El modelo de Cox extendido

Al modelo de Cox en el que se incluyen variables dependientes del tiempo se le denomina modelo de Cox extendido[31]. La expresión del modelo de Cox extendido es

$$h(t; z(t)) = h_0(t) \times e^{z(t)\beta}$$

[31] Sobre este modelo ya hemos hablado de forma sucinta en el apartado 3.2.2.3 en el que el modelo de Cox extendido nos permitió evaluar la hipótesis de proporcionalidad.

donde $z(t)$ es un vector de variables explicativas que incluye tanto variables no dependientes del tiempo x como variables dependientes del tiempo $y(t)$. Otra forma alternativa de expresar, por lo tanto, este modelo es

$$h(t; z(t)) = h_0(t) \times e^{x\beta + y(t)\gamma}$$

que nos permite resaltar algunas de las implicaciones de este modelo: en primer lugar, el efecto de $z(t)$ sobre la función de riesgo en t depende tan sólo de los valores de las variables en ese mismo instante t y no de valores futuros o pasados y, en segundo lugar, el efecto es constante en el tiempo como queda palpable al especificar el parámetro que recoge dicho efecto como un parámetro constante. En ocasiones pueden existir razones para creer que hay un cierto retardo entre el cambio que sufre una variable y el efecto que ese cambio produce sobre la función de riesgo. Esto no supone ninguna dificultad pues el modelo de Cox extendido se puede reformular para incluir estos efectos retardados. Siendo r el indicador del número de retardos, la especificación del modelo de Cox extendido con variables retardadas es,

$$h(t; z(t)) = h_0(t) \times e^{x\beta + y(t-r)\gamma}$$

Resaltamos que este modelo ya no verifica la hipótesis de proporcionalidad, lo que significa que el cociente de las funciones de riesgo de dos individuos dados con características recogidas por los vectores $z_1(t)$ y $z_2(t)$ ya no será constante sino que va a depender de t como se muestra en la siguiente expresión:

$$\frac{h(t; z_1(t))}{h(t; z_2(t))} = e^{(z_1(t) - z_2(t))\beta}$$

ya que aunque se sigue anulando la función de riesgo base, la parte que recoge el efecto de las variables ahora depende también de t[32]. Además la interpretación de los coeficientes se ve también alterada, ya que los parámetros asociados a variables dependientes del tiempo recogen efectos medios de esas variables cuando se consideran todos los valores de t para los que se ha medido la variable.

La estimación del modelo requiere que los datos sean expandidos para poder incorporar los diferentes valores de cada variable, de forma muy similar a lo que se

[32]Esta doble dependencia del tiempo dificulta la distinción entre el efecto de una variable que varía en el tiempo y la dependencia de la duración. En general, se requiere que haya mucha variación en esas variables dependientes del tiempo entre los individuos.

hace en los modelos de duración en tiempo discreto. En este caso, es necesario dividir el tiempo en subperiodos para los cuales la variable se mantiene constante. Esto es, se construyen intervalos de duración para los cuales $z(t)$ no varía y de tal manera que tengamos tantas observaciones correspondientes a un mismo individuo como valores toma la variable $z(t)$.

Aquí la mayor complicación se debe a la información disponible sobre el comportamiento de estas variables dependendientes del tiempo. En concreto, estamos refiriéndonos a la frecuencia con que han sido recogidos los datos. En general, las variables suelen medirse a intervalos de tiempo regulares que no tienen porque coincidir con los tiempos en los que se ha producido una observación de t. Principalmente, el problema surge cuando el tiempo de ocurrencia del suceso se mide de forma más precisa que las variables dependientes del tiempo. Este sería el caso de un estudio sobre la duración del desempleo en el que la variable dependiente ha sido medida en días y en el que la variable explicativa que nos indica las condiciones de mercado en las que se mueve el individuo que está buscando empleo, por ejemplo la tasa de paro local, está medida en meses. En estas situaciones, una posible solución es estimar los valores de esas variables explicativas para esos tiempos intermedios bien utilizando el valor observado más próximo o bien interpolando.

3.5 Heterogeneidad no observada

3.5.1 Introducción

En los modelos expuestos hasta el momento se ha supuesto que disponemos de información exhaustiva sobre las características que diferencian a los individuos en cuanto a su duración, esto es, el vector X contiene todos aquellos factores que han tenido alguna influencia sobre la duración y que hacían que la población no fuera homogénea en cuanto a su comportamiento. Ahora bien, como señala Lancaster (1990), no siempre es posible especificar correctamente un modelo de duración ya que, por un lado, ciertas variables de ese vector no son observables o no han podido ser observadas y, por otro lado, las variables pueden presentar errores en su medición. Por lo tanto, el término de *heterogeneidad no observada* en este contexto se refiere a la existencia de diferencias en la distribución de la variable duración entre los individuos que no han podido ser controladas a través de las variables explicativas incluidas en el modelo.

3.5.2 Especificación

La propuesta más aceptada en la literatura para solucionar este problema ha consistido en la consideración de un nuevo factor que recoge todo aquello antes comentado. La materialización de esta propuesta supone la inclusión de una variable aleatoria positiva no observable cuya media es igual a uno y tiene una varianza finita desconocida. La varianza de esta variable es entonces la que nos va a permitir medir el grado de heterogeneidad no controlada por el modelo especificado[33]. A este tipo de modelos de duración, que incluyen un factor de heterogeneidad no observada, se les denomina modelos de duración mixtos porque, como veremos, son el resultado de mezclar dos distribuciones, esto es, una mixtura. En general, el factor de heterogeneidad varía para cada individuo aunque en situaciones específicas se puede considerar que es común para los individuos pertenecientes a un mismo grupo, por ejemplo, cuando las observaciones corresponden a miembros de una misma familia[34].

La incorporación de ese factor en cualquiera de los modelos considerados hasta ahora, es relativamente sencilla. En concreto, si partimos del modelo de riesgos proporcionales, que se caracteriza porque las variables -sean observadas o no- tienen un efecto multiplicativo, y de la expresión de la función de riesgo,

$$h(t;z) = h_0(t)e^{z\beta}$$

donde z, la matriz de observaciones de las variables, se puede particionar de la siguiente forma $z = (x \quad y)$, de modo que x contenga las variables observadas e y las no observadas, obteniendo,

$$h(t;z) = h_0(t)e^{x\beta}e^{y\gamma}$$

y, por lo tanto,

$$h(t;x,v) = h_0(t)e^{x\beta}v$$

[33] Algunos autores continuando con el trabajo pionero de Heckman y Singer (1984) han optado por una especificación no paramétrica de la heterogeneidad no observada.

[34] El término en inglés que reciben estos últimos es el de modelos mixtos con un "*shared frailty*". El vocablo "*frailty*" fue acuñado por primera vez en un estudio sobre la mortalidad realizado por Vaupel *et al.* (1979) y con él se trata de recoger la idea que subyace en estos modelos, esto es, la existencia de individuos que son más frágiles que otros.

que es la expresión de un modelo de duración mixto con riesgos proporcionales. El efecto multiplicativo del factor de heterogeneidad supone que en el caso de que exista dependencia positiva, ésta se va a ver agrandada para aquellos individuos que presentan una característica no observada con un efecto igualmente positivo sobre el riesgo.

En este modelo subyacen dos hipótesis con importantes implicaciones en la estimación. En primer lugar, se asume que existe independencia entre las variables explicativas observadas y las no observadas. En segundo lugar, el factor v aunque varía para cada individuo es constante en el tiempo. Esta última hipótesis supone asumir que lo que no es observado para un determinado individuo en un periodo tampoco se observa en cualquier otro periodo.

3.5.3 Efectos de la no inclusión de la heterogeneidad no observada

Las consecuencias de no tener en cuenta la *heterogeneidad no observada* han sido puestas de relieve por diferentes autores desde un punto de vista teórico (Heckman y Singer, 1984) y desde un punto de vista empírico (Lancaster, 1979). En términos generales, este error de especificación va a afectar a la dependencia de la duración y al efecto de las variables explicativas sobre la duración.

Dado que el término v no es observable, los datos que se observan son generados por una distribución mixta, siendo la relación entre esta distribución mixta $f_m(t;x)$ y la verdadera distribución no observable $f(t;x,v)$,

$$f_m(t;x) = \int_0^\infty f(t;x,v)\, dF(v) = E_v[f(t;x,v)]$$

La existencia de dependencia de la duración viene determinada por el comportamiento de la función de riesgo en el tiempo. Si tomamos la función de riesgo de la distribución mixta

$$h_m(t;x) = E_v[h(t;x,v)] = h(t;x)E[V/T \geq t]$$

y aplicamos neperianos y derivamos, obtenemos que

$$\frac{d}{dt}\ln h_m(t;x) = \frac{d}{dt}\ln h_0(t) - h_0(t)\frac{var(V/T \geq t)}{E[V/T \geq t]}$$

donde el segundo término de esta expresión es siempre una cantidad positiva por lo que se concluye que en cualquier caso hay una sobreestimación (infraestimación) de la dependencia de la duración negativa (positiva). Esto es, en el caso de heterogeneidad

no observada, los individuos con los valores más altos del factor v abandonan el estado, por término medio, más rápidamente, y los individuos que restan son los que presentan los valores más bajos de v, de ahí, esta tendencia a observar dependencia de la duración negativa.

En el trabajo de Lancaster (1979), pionero en la utilización de los modelos de riesgos proporcionales mixtos, se muestra como el parámetro de la distribución Weibull va aproximándose a uno a medida que se incluyen más regresores[35]. En palabras de Lancaster, el parámetro actúa como una varianza residual en la medida que cuantos más regresores significativos se incluyen, menor es su valor.

Respecto a cómo se ven afectados los parámetros cuando no se tiene en cuenta la heterogeneidad, se demuestra que

$$\frac{d}{dx_k}\ln h_m(t;x) = \beta_k\left[1 - e^{x\beta}H_0(t)\frac{var(V/T \geq t)}{E[V/T \geq t]}\right]$$

expresión que de nuevo nos lleva a hablar de una infraestimación de los valores de los parámetros β_k ya que la expresión que aparece entre corchetes es siempre negativa. Además, este efecto, como se deduce de la ecuación anterior, no es constante en el tiempo.

Por lo tanto, ante los graves problemas que acarrea un error de especificación en los modelos de duración, es lógico que en la literatura se haya planteado la realización de algún contraste que nos determine la existencia de heterogeneidad no observada. En el caso de que se haya optado por una modelización paramétrica del factor de heterogeneidad, se recurre a un constraste de razón de verosimilitud en el que se contrasta que la varianza de este factor es igual cero. La aceptación de la hipótesis nula nos lleva a rechazar la necesidad de incluir un factor de estas características en el modelo.

3.5.4 Estimación

En cuanto a la estimación de estos modelos mixtos la cuestión está aún en desarrollo, habiéndose alcanzado ciertos resultados definitivos en el caso de los modelos paramétricos, esto es, en aquellos modelos en los que la función de riesgo base es especificada paramétricamente.

[35]Recuérdese que cuando el parámetro es igual a uno estaríamos en el caso de una distribución exponencial, única distribución que no muestra dependencia de la duración.

En este caso, a la hora de estimar hemos de tener en cuenta que el factor v incluido en el modelo es una variable no observable y por lo tanto no se dispone de información sobre los valores que toma para cada individuo v_i. De ahí que sea necesario especificar una distribución para este factor de tal manera que quede caracterizado por un número finito de parámetros que sí se pueden estimar. Una vez seleccionada la distribución, la maximización de la función de verosimilitud nos lleva a obtener junto con los parámetros de interés, la varianza del factor de heterogeneidad (θ),

$$\ln L(\beta, \theta) = \sum c_i \ln f_m(t_i; x) + (1 - c_i) \ln S_m(t_i; x)$$

Por motivos de operatividad, las distribuciones candidatas han de proporcionar una forma sencilla de la función de supervivencia. Esto ha hecho que el número de distribuciones que ha sido utilizado en la práctica sea muy reducido. En concreto, podemos hablar de la distribución gamma propuesta por Lancaster (1979) y la distribución gaussiana inversa. Las funciones de riesgo de estas dos distribuciones no son muy distintas,

$$h_m(t; x) = h(t; x)[1 - \theta \ln S(t; x)]^{-1}$$

para la distribución gamma y, para la distribución gaussiana inversa

$$h_m(t; x) = h(t; x)(1 - 2\theta \ln S(t; x))^{-1/2}$$

Las aplicaciones prácticas de estos modelos mixtos muestran en general que los resultados dependen en gran medida de la forma funcional elegida para parametrizar la función de riesgo base. Esto ha llevado a plantear modelos mixtos en los que la forma de la función de riesgo sea más flexible como es el caso del modelo de Cox. Sin embargo, Meyer (1990) muestra que en la práctica cuando se especifica no paramétricamente la función de riesgo, como es el caso del modelo de Cox, la inclusión de un factor de heterogeneidad, en general, sólo complica de forma considerable la estimación del modelo sin apenas modificar los resultados en comparación al modelo sin incluir la heterogeneidad no observada.