Programa de Especialización en Econometría Aplicada
Centro de Formaciòn Continua -UNI
Modelos de Supervivencia
Clase 4

Edinson Tolentino
MSc Economics
email: edinson.tolentino@gmail.com

Twitter: @edutoleraymondi

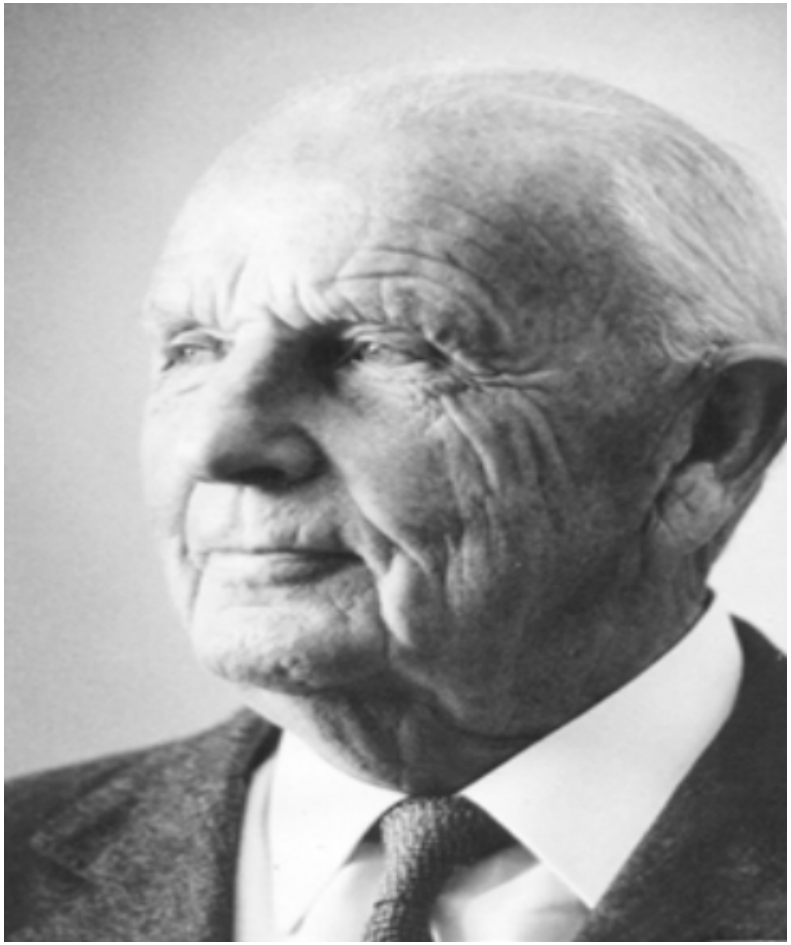
Universidad Nacional de Ingeneria

17 de junio de 2023

Contenido

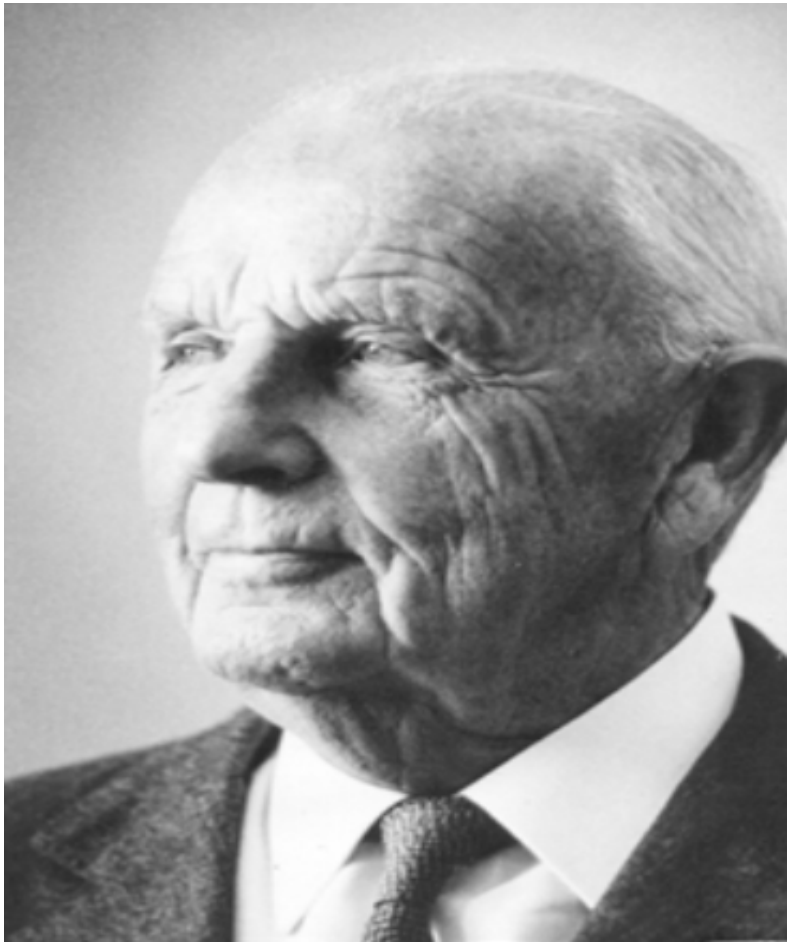
Introducción

Introducción



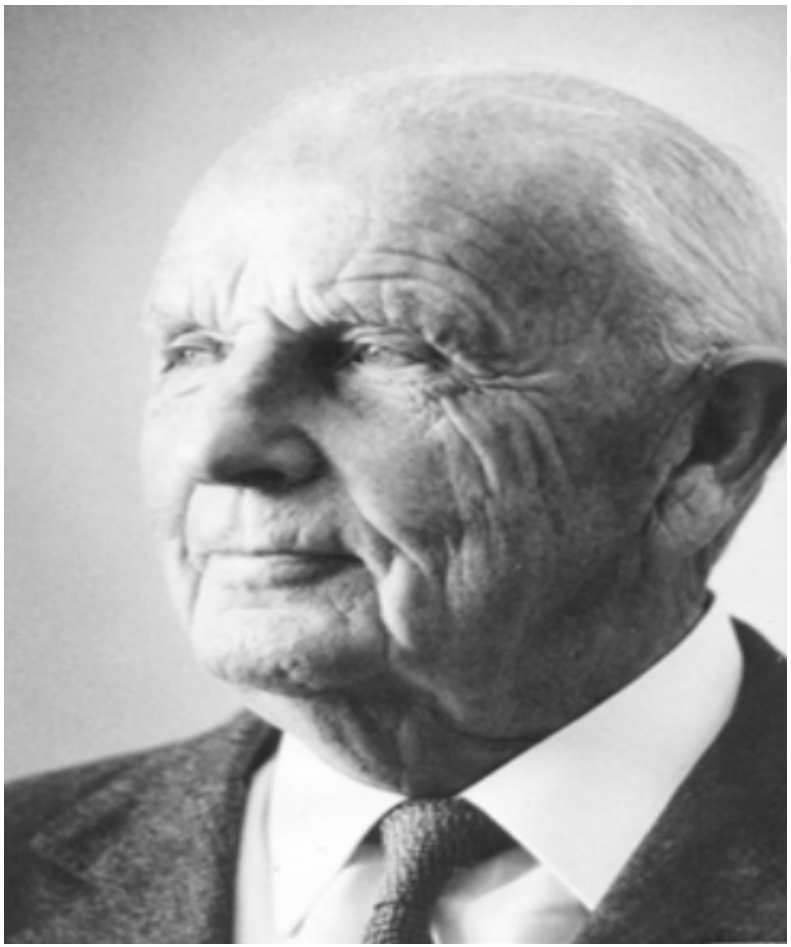
- Ernst Waloddi Weibull (1887-1979)

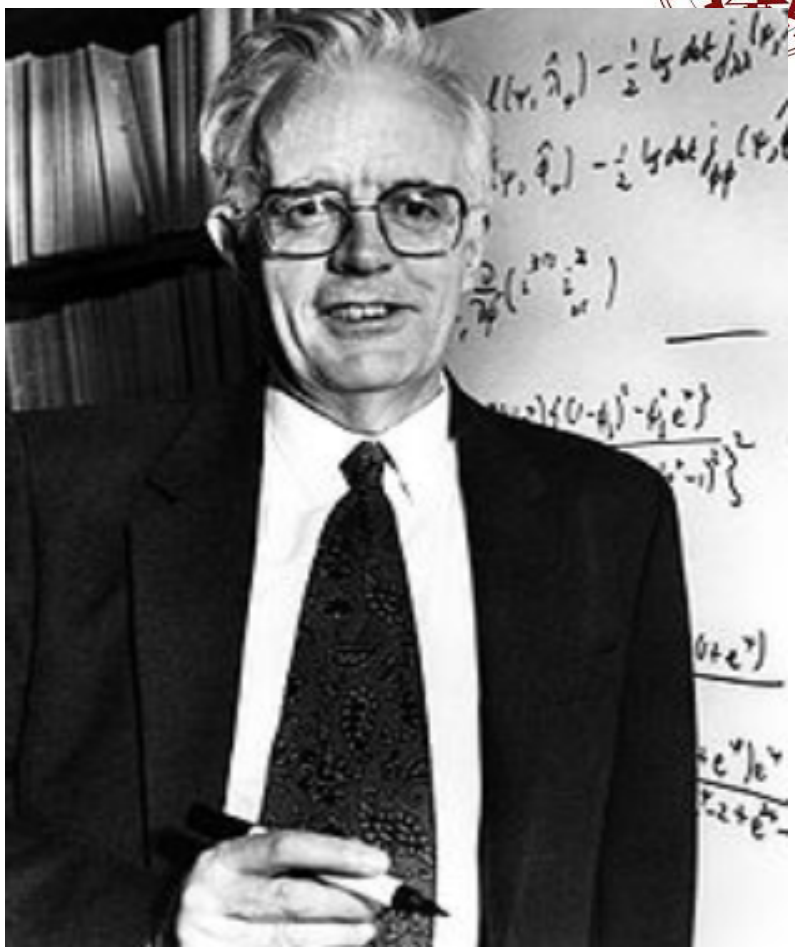
Introducción



- Ernst Waloddi Weibull (1887-1979)
- El modelo de Proporciones Weibull

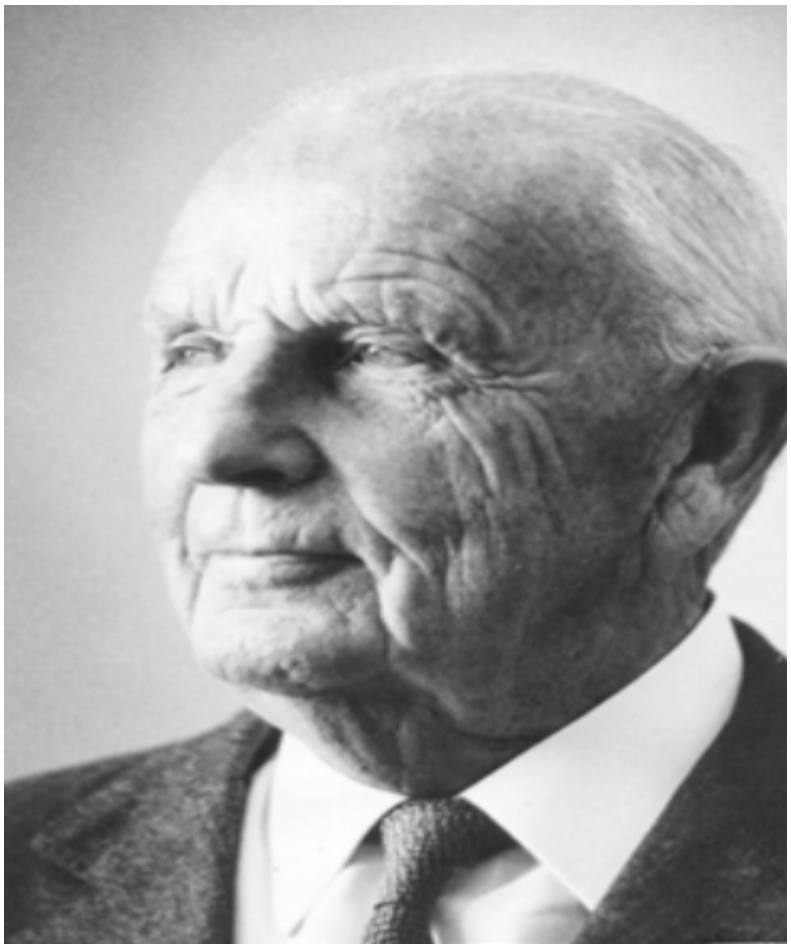
Introducción

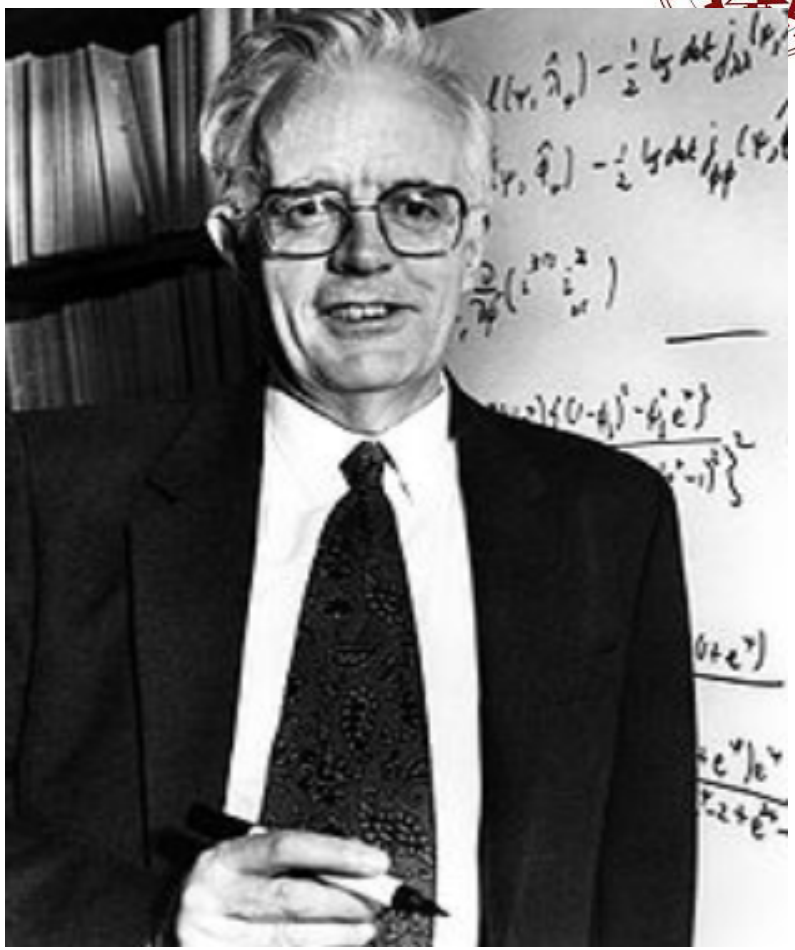




- Ernst Waloddi Weibull (1887-1979)
- El modelo de Proporciones Weibull
- David Roxbee Cox (1924 - presente)

Introducción





- Ernst Waloddi Weibull (1887-1979)
- El modelo de Proporciones Weibull

- David Roxbee Cox (1924 - presente)
- El modelo de Proporciones de Cox

Introducción

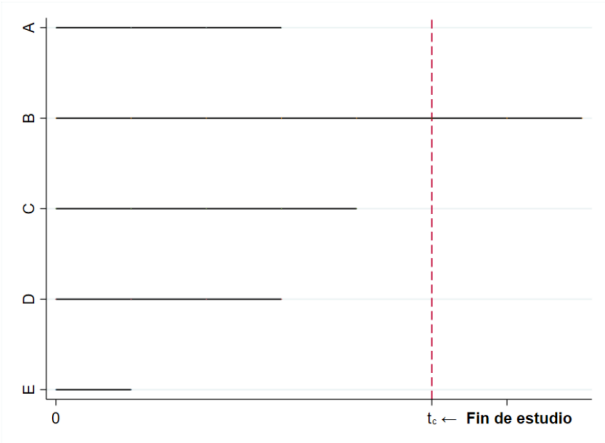
- Hasta el momento, la dimensión "tiempo" ha sido excluido del análisis de microdatos. Ahora extenderemos nuestra visión hacia modelos donde la variable dependiente esta directamente relacionado al "tiempo", pero manteniendo la idea de "cross section" (y con ello, la independiencia de observaciones).
- En sencillo, los modelos de duración analizan el tiempo que ha transcurrido entre la ocurrencia de dos eventos particulares. Algunos ejemplos:
  - Semanas de desempleo.
  - Tiempo que tarda un delincuente en reincidir.
  - Meses/trimestres/años hasta que una empresa salga del mercado.
  - Duración de huelgas o matrimonios
- El análisis de duración tiene origen en el denominado "Análisis de sobrevivencia", donde la duración de interés es el tiempo de vida de un sujeto. ASí, el interés recae en determinar cómo varios tratamientos o las características afectan los tiempo de vida.
- En Ciencias Sociales, el interés se centra en cualquier situación en la cual un individuo (hogar, empresa, etc.) comienza en un estado inicial y se observa la salida de este estado o es censurado.

Falla y Tiempo de Falla

- La "Falla" (failure) es un evento puntual que define el cambio del estado actual. Ejemplo: Desempleo $\rightarrow$ Empleado (failure).
- El "Tiempo de Falla" (Failure time) se refiere al tiempo transcurrido hasta la ocurrencia de la Falla.
- Existen tres condiciones para determinar el Failure Time precisamente:
  1. El tiempo de origen debe ser definido sin ambiguedades. Sin embargo, no necesariamente todas las unidades de análisis deben iniciar al mismo tiempo.
  2. La escala de medición del tiempo debe ser uniforme (horas, meses, etc.).
  3. Las condiciones y casos de Falla deben estar determinados claramente. Ejemplo: Desempleo $\rightarrow$ Empleado (No capacitación.)

Censura

- Un problema inherente al análisis de supervivencia es la censura de observaciones puesto que la recolección de datos final se realiza cuando el proceso de interés continúa.
- No tomar en cuenta la censura de datos trae como consecuencia estimar coeficientes sesgados. Así, es necesario considerar la censura en la estimación.



Formalmente, podemos escibir la duración observada como: $t_i = \min(t_i^*, t_c)$
Donde:

- t_i es la duración o tiempo de falla observada.
- t_i^* es la duración o tiempo de falla real.
- t_c es el momento de la censura

Entonces:
Si $t_i^* \leq t_c \Rightarrow y_i = 1 \rightarrow$ Observación no censurada.
Si $t_i^* > t_c \Rightarrow y_i = 1 \rightarrow$ Observación censurada

Función de Supervivencia

- Sea $T \geq 0$ la variable de duración (tiempo en que una persona deja el estado inicial) y t denota un valor particular de T. Podemos definir la **función de distribución acumulada** de T como:

$$F(t) = Pr(T \leq t) \; ; \; t \leq 0$$

- $F(t)$ (Failure Function) indica la probabilidad de que el evento dure como máximo hasta t. Así, la **función de supervivencia** (survivor function) esta definida como:

$$S(t) = Pr(T > t) = 1 - F(t)$$

  Representa la probabilidad de sobrevivir pasado el momento t (en otros palabras, indica la probabilidad de que el evento dure por lo menos hasta t).

- Aimismo, la función de densidad de T se denota por:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

Función de Riesgo

- En muchas ocasiones estaremos interesados en conocer la probabilidad de dejar el estado inicial en el intervalo $(t, t+h)$ dada la sobrevivencia hasta el momento t.

$$Pr(t \leq T < t+h \,|\, T \geq t)$$

- A partir de ello, la **función de riesgo (hazard function)** se define como:

$$\lambda(t) = \lim_{h \to 0} \frac{Pr(t \leq T < t+h \,|\, T \geq t)}{h}$$

para cada t, $\lambda(t)$ es la tasa instantánea de salida por unidad de tiempo.

- Si T es la duración de una microempresa en el mercado, medida en años, entonces $\lambda(10)$ es aproximadamente la probabilidad de dejar el mercado entre el año 10 y 11 condicional en haber estado en el mercado hasta el año 10.
- En otras palabras, $\lambda(t)$ indica la probabilidad de dejar el mercado durante el año 11, dado que la microempresa ha permanecido en el mercado los 10 años anteriores.

Función de Riesgo

- Ahora, de estadística, sabemos que $f(x/y) = \frac{f(x,y)}{f(y)}$ por lo que:

$$Pr(t \leq T < t+h | T \geq t) = \frac{Pr(t \leq T < t+h; T \geq t)}{Pr(T \geq t)}$$

Además:

$$Pr(\frac{A}{B}) = \frac{Pr(A \cap B)}{Pr(B)} \quad \text{y} \quad Pr(\frac{B}{A}) = \frac{Pr(A \cap B)}{Pr(A)}$$

Así, definimos $A = t \leq T < t+h$ y $B = T \geq t$ para expresar:

$$Pr(t \leq T < t+h | T \geq t) = \frac{Pr(t \leq T < t+h \cap T \geq t)}{Pr(T \geq t)} = \frac{Pr(t \leq T < t+h)}{Pr(T \geq t)}$$

Función de Riesgo

Entonces:

$$\begin{aligned}
\lambda(t) &= \lim_{h\to 0} \frac{Pr(t \leq T < t+h | T \geq t)}{h} \\
&= \lim_{h\to 0} \frac{Pr(t \leq T < t+h)}{Pr(|T \geq t).h} \\
&= \lim_{h\to 0} \frac{F(t+h) - F(t)}{[1-F(t)].h} \\
&= \lim_{h\to 0} \frac{F(t+h) - F(t)}{h} \frac{1}{S(t)} \\
\lambda(t) &= \frac{f(t)}{S(t)}
\end{aligned}$$

- La función de riesgo puede ser definida como el cociente de la densidad sobre la función de supervivencia.

Función de Riesgo

Puesto que $f(t) = \frac{-dS(t)}{dt}$:

$$\lambda(t) = \frac{f(t)}{S(t)}$$
$$= \frac{-dS(t)/dt}{S(t)}$$
$$\lambda(t) = \frac{-d \ln S(t)}{dt}$$

- Otro concepto importante es la **integral de la función hazard**:

$$H(t) = \int_0^t \lambda(s)ds = \int_0^t -\frac{d \ln S(s)}{ds} ds = -\ln S(t) + \ln S(0)$$
$$H(t) = -\ln S(t)$$

Función de Riesgo

Puesto que $S(t) = 1 - F(t)$:

$$H(t) = -\ln S(t)$$
$$e^{H(t)} = -S(t)$$
$$e^{H(t)} = -[1 - F(t)]$$
$$F(t) = 1 - exp[\int_0^t \lambda(s)ds]$$

Diferenciando:

$$f(t) = \lambda(t).exp[\int_0^t \lambda(s)ds]$$

- Este resultado muestra como es posible reconstruir la función de supervivencia $S(t)$, e inmediatamente $F(t)$, a partir de la función de riesgo.
- Así, podemos construir el modelo de duración basado exclusivamente en la función de riesgo, puesto que caracteriza muy bien la distribución de una variable aleatoria continua (T).

Modelos paramétricos

- Antes de asignar una forma funcional a la función de riesgo es importante ver su comportamiento respecto al tiempo.
- La naturaleza de la relación entre la función de riesgo y la duración se le conoce como **dependencia de duración** y es posible encontrar 3 casos:
  1. **Neutra** $(\partial h/\partial t = 0)$: El ratio de riesgo es constante o no reacciona al periodo de duración. En el ambito de la economía laboral, esto implicaría que la probabilidad de encontrar empleo no depende del tiempo en que la persona se encuentre desempleado.
  2. **Positiva** $(\partial h/\partial t > 0)$: La probabilidad de cambiar de estado desempleo $\rightarrow$ empleado aumenta en la medida que mayor es la duración en el primer estado.
  3. **Negativa** $(\partial h/\partial t < 0)$: La probabilidad de cambiar de estado disminuye con la duración.
- Existen **distribuciones** muy útiles para el modelamiento de la duración. El uso de distribuciones especificas para el modelamiento de la duración es llamado efoque paramétrico.

Modelos paramétricos

Distribuciones para modelos de duración

Distribución	F(t)	S(t)	h(t)	H(t)
Exponencial	$1-e^{-\lambda t}$	$e^{-\lambda t}$	λ	λt
Weibull	$1-e^{(-\lambda t)^{\alpha}}$	$e^{-\lambda t^{\alpha}}$	$\alpha\lambda t^{\alpha-1}$	λt^{α}
Gompertz		$e^{-\frac{\gamma}{\alpha}(e^{\alpha t-1})}$	$\gamma e^{\alpha t}$	$\frac{\gamma}{\alpha}(e^{\alpha t}-1)$
Log-Logística	$\frac{(\lambda t)^{\alpha}}{1+(\lambda t)^{\alpha}}$	$\frac{1}{1+(\lambda t)^{\alpha}}$	$\frac{\alpha\lambda^{\alpha}t^{\alpha-1}}{1+(\lambda t)^{\alpha}}$	$\log(1+\lambda t^{\alpha})$
Log-Normal		$1-\phi(\frac{\ln t-\mu}{\sigma})$	$\frac{e^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}}{t\sigma\sqrt{2\pi}[1-\Phi(\frac{\ln t-\mu}{\sigma})]}$	$-\log S(t)$

Modelos paramétricos

A continuación analizaremos las funciones exponencial y Weibull.

Distribución Exponencial

- Supongamos que T sigue una distribución exponencial

$$F(t) = 1 - e^{-\lambda t}$$
$$S(t) = e^{-\lambda t}$$
$$h(t) = \lambda$$

- Esta distribución tiene una función de riesgo constante y es conocida como **memoryless**: la probabilidad instantánea de que el evento concluya, condicional al pasado de la misma, no varía en el tiempo.
- El pasado no contribuye a aumentar o disminuir esta probabilidad condicional. También, se dice que con esta distribución no hay dependencia de la duración, es decir, la función de riesgo es independiente del tiempo.

Modelos paramétricos

Distribución Weibull

- De acuerdo a esta distribución

$$F(t) = 1 - e^{(-\lambda t)^{\alpha}}$$
$$S(t) = e^{-\lambda t}$$
$$h(t) = \alpha \lambda t^{\alpha - 1}$$

- La distribución Weibull permite definir una función de riesgo monot+onicamente creciente o decreciente, dependiendo de sus parámetros.
- λ es el parámetro de escala y α es el parámetro de forma. α define si $h(t)$ es monotónicamente creciente o decreciente con el tiempo.
  1. Si $\alpha = 1 \Rightarrow \partial h/\partial t = 0 \Rightarrow$ Distribución Exponencial
  2. Si $\alpha > 1 \Rightarrow \partial h/\partial t > 0$
  3. Si $\alpha < 1 \Rightarrow \partial h/\partial t < 0$
- La distribu Weibull permitirá modelar procesos que sean solo crecientes o solo decrecientes.

Estimación

- Los parámetros α y λ de las distribuciones propuestas, para modelar la duración, pueden ser estimadas mediante máxima verosimilitud.
- Consideremos una muestra de n individuos para la cual observamos un conjunto de personas (empresas) las cuales han completado su duración y otras para las cuales seguimos observando su situación inicial.
- Sea t_i la duración observada para el individuo i. Definiremos:

$$t_i = \text{mín}(t_i^*, t_c)$$

Donde t_i^* es la duración real y t_c el momento de censura. Ahora, definimos la probabilidad que t_i sea censurada:

$$Pr(t_i^* \geq t_c | x_i) = 1 - F(t_c | x_i, \beta)$$

Donde $F(t_c | x_i, \beta)$ es la función de distribución acumulada de t_i^*.

Estimación

- Además, sea d_i un indicador de censura ($d_i = 1$ si no hay censura y $d_i = 0$ o si hay censura), la verosimilitud condicional para la observación i puede escribirse como:

$$f(t_i|x_i, \beta)^{d_i}[1 - F(t_i|x_i, \beta)]^{1-d_i}$$

Desde la cual podemos obtener la función log $-$likelihood, asumiendo una muestra de n individuos:

$$\log L(\beta, t, x) = \sum_{i=1}^{n} d_i \log f(t_i|x_i, \beta) + (1 - d_i) \log[1 - F(t_i|x_i, \beta)]$$

Introducción

- Las variables exógenas o regresores se denominan covariables y puede ser:
  1. Covariables que no varían en el tiempo (**time-invariant**), como pueden ser el géneroy la raza.
  2. Covariables que varían en el tiempo (**time-varying**), como la edad o años de educación.
- Cuando estas variables no cambian en el tiempo, se define el riesgo (y todas las otras características de T) condicional en las covariables:

$$\lambda(t,x) = \lim_{h\to 0} \frac{Pr(t \leq T < t+h | T \geq t; x)}{h}$$

donde x es un vector de explicativas. Todas las formulas anteriores se mantienen considerando que tanto la función de distribución acumulada como la densidad son condicionales en x. Por ejemplo:

$$\lambda(t,x) = \frac{f(t|x)}{1-F(t|x)}$$

donde $f(t|x)$ es la densidad de T condicional en x.

- La adición de variables explicativas en los modelos de duración no es trivial, y dependen de la interpretación deseada. A continuación se presentan algunas estrategpias comunmente utilizadas.

Modelo de Riesgo Proporcional

- Los modelos de duración con distribución Exponencial, Weibull y Gompertz puede ser formuladas como Modelos de Riesgo Proporcional (Proportional hazard (PH) models).
- En esyta especificación, las variables explicativas afectan directamente a la función de riesgo en forma proporcional. En este caso, el riesgo proporcional puede escribirse como:

$$\lambda(t, x) = \kappa(x)\lambda_0(t)$$

  donde $\kappa(\cdot)$ es una función nonegativa de x y $\lambda_0(t)$ se denomina riesgo base.
- El riesgo base es común a todas las unidades de la población, las funciones de riesgo proporcional difieren principalmente basados en una función $\kappa(x)$ de variables observables.

Modelo de Riesgo Proporcional

- Por ejemplo, asumimos $\kappa(x) = e^{x'\beta}$ (función exponencial).

$$\lambda(t,x) = e^{x'\beta}.\lambda_0(t)$$
$$\ln\lambda(t,x) = x'\beta + \ln\lambda_0(t)$$
$$\frac{\partial\lambda(t,x)}{\partial x_k} = \beta_k \quad \text{ó} \quad \frac{\partial\lambda(t,x)}{\partial x_k} = \lambda_0(t)\beta_k e^{x'\beta}$$

- Así, β_k otorga el cambio proporcional en la función de riesgo que resultad de un cambio marginal en la k-esíma variable explicativa. Es decir, los coeficientes estimados tienen una interpretación de semielasticidad de la función de riesgo con respecto a las covariables.
- Cuando x_k es una variables binaria, la interpretación adecuada es la siguiente. Supongamos que x y x^* difieren solo en la k-esíma variable explicativa, la cual es binaria

$$\Rightarrow \frac{h(t|x^*)}{h(t|x)} = e^{\beta_k}$$

- Este resultado proporciona el riesgo de que el evento ocurra si el individuo pertenece a cierto grupo (indicado por x_k) en relación al riesgo de que el evento ocurra si el individuo no pertenece a dicho grupo.

Modelo de Riesgo Acelerado

- Las funciones de riesgo que siguen funciones log $-$normal o log $-$logística no pueden ser clasificados como PH models. Alternativamente , ambos modelos pueden ser resumidos en el modelo de riesgo acelerado (accelerated failure time (AFT) models).
- Este modelo incorpora las variables explicativas a través de la función de supervivencia. ASí, el logaritmo de t_i es expresado en función a covariables y un término de error aditivo.

$$\ln t_i = x_i'\beta + \mu_i$$

- Supuesto sobre la distribución de μ_i determinan la distribución de t_i y por lo tanto de la forma funcional de la función hazard.
- Así, si μ_i sigue una distribución normal (o logística), obtendremos el modelo log $-$normal (o log $-$logística). Además, podemos obtener el modelo Weibull (y con ello el exponencial) si μ_i asume una distribución xtreme $-$ value.

Modelo de Riesgo Acelerado

- Para conocer como los AFT models aceleran el tiempo de falla, reescribimos la función anterior como:

$$\ln \psi t_i = \mu_i$$

donde $\psi = \psi(x, \beta)$. Notese que en este caso, el efecto de las variables explicativas es alterar la escala temporal del evento.

- La función de supervivencia es, por definición, una función monótona decreciente de t. Entonces, $\psi(x, \beta)$ es un factor de aceleración que indica como las variables explicativas afectan a la escala temporal.
- En los AFT models, la interpretación de los coeficientes es:

$$\frac{\partial \ln t_i}{\partial x_k} = \beta_k \quad \text{o} \quad \beta_k \times 100\,\% \quad \text{o} \quad [e^{\beta_k} - 1] \times 100\,\%$$

Modelos de Riesgo Proporcional de COX

- Dado el modelo de Riesgo Proporcional, COX (1972) sugiere un método de máxima verosimilitud parcial para su estimación.

$$\lambda(t, x) = \kappa(\beta' x)\lambda_o(t)$$

- La diferencia clave entre la propuesta de COX (1972) y el resto de modelos es la formulación del riesgo base $(\lambda_0(t))$.
- Así, en COX (1972), la forma funcional para el riesgo base es no especificada y por lo tanto adquiere una forma flexible.
- A partir de ello, el modelo de COX es conocido como **modelo de Riesgo Proporcional Semiparamétrico** porque este solo hace un supuesto paramétrico sobre cómo las variables explicativas impactan al riesgo (distribución exponencial, ...) y no hace un supuesto distribucional sobre la naturaleza del riesgo base.
- El riesgo base $(\lambda_0(t))$ puede ser interpretado como un intercepto que varía con el tiempo.

Modelo de COX (1972)

- COX (1972) demostró que los parámetros β puede ser estimados sin especificar ninguna forma funcional para $\lambda_0(t)$. Por lo tanto, se puede conocer el efecto de x_j sin necesidad de tomar una forma funcional para $\lambda_0(t)$.
- Sin embargo, tiene un costo y es que no se podrá conocer la naturaleza de la dependencia de la duración $(\partial\lambda/\partial t)$ y también podría generarse una complicación cuando se presentan varias fallas a la vez.

Ventajas

1. Proporciona una estimación de los β sin especificar el riesgo base. Lo cual es visto como una forma más flexible de modelar respecto a los modelos paramétricos.
2. El inconveniente del uso de modelos paramétricos es que si el modelo esta mal especificado a traves de un supuesto distribucional inapropiado, la estimación otorga parámetros inconsistentes.
3. Sin embargo, un modelo semi-paramétrico como el de COX, proporciona algún grado de protección frente a este tipo d emalas especificaciones si su procedencia redica en el riesgo base.

Modelo de COX (1972) - Supuesto

- El supuesto incorporado asume que el efecto de los covariantes sobre el ratio hazard es constante en el periodo de análisis. Si este supuesto es violado entonces las estimaciones inconsistentes e ineficientes.
- Es posible evaluar, gráfica y estadisticamente, el supuesto de riesgo proporcional dentro del modelo de COX. Para ello, necesitamos los residuos de **Schoenfeld**.

Forma gráfica

- Schoenfeld (1980) propone calcular el residuo como la diferencia entre el valor observado del predictor asociado al parámetro del modelo y el valor esperado del predictor para los casos con el riesgo establecido en el momento del evento observado.

$$r_k(\hat{\beta}) = x_{(k)} - E(X_{(k)})$$

- Así, para un predictor x, los gráficos de los residuos de Schoenfeld respecto del tiempo mostraría una linea horizontal en 0 si se cumple el supuesto de proporcionalidad de los riesgos.

Test formal

- Se realiza una regresión auxiliar donde la variable dependiente son los residuos de Schoenfeld y la covariable es el tiempo. Así, se realiza un test de wald para ver la significancia del coeficiente estimado. Si el coeficiente es estadísticamente igual a cero entonces el supuesto de riesgo proporcional se cumple.

Métodos no paramétricos

- Los economistas están generalmente interesados en modelar las funciones de riesgo. Así, siempre es conveniente gráficar la función hazard empírica. Para ello, procedimientos no paramétrico como Kaplan-Meier Function, Nelson-Aalen y tablas de vida son generalmente utilizado:

Kaplan-Meier

$$\hat{S}(t) = 1 - \hat{F}(t)$$

$\Rightarrow$ estimar $\theta = F(t) = Pr(T \leq t)$ considerando censura:

$$\hat{\theta}_i = \frac{n_i - d_i}{n_i}$$

donde:

- $n_i = \#$ de individuos en situación de riesgo al comienzo del i-esímo intervalo.
- $d_i = \#$ de individuos que experimentaron el evento.

Así:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i}$$

Métodos no paramétricos

Nelson-Aalen

$$\hat{S}(t) = 1 - \hat{F}(t)$$

$\Rightarrow$ estimar $\theta = F(t) = Pr(T \leq t)$ para cada t

$$\hat{\theta} = \frac{\#observaciones \leq t}{n_i}$$

Lifetable

El estimador de la función de riesgo acumulado es:

$$\hat{H}(t_i) = \sum_{i:t_i<t} \frac{d_i}{n_i}$$

$$\hat{S}(t_i) = e^{\hat{H}(t)}$$

Implementación en Stata

Antes de estimar cualquier modelo de duración, primero tenemos que decirle a Stata que vamos a realizar un análisis de supervivencia. Para ello usamos el comando **stset** (Survivel Time Set).

Syntax

stset `timevar [,failure(failvar)]`

donde `timevar` es la variable duración y `failvar` describe los eventos. Luego se sugiere utilizar el comando **stsum** para resúmenes estadísticos de la data.
Seguido, es importante observar la función de supervivencia. Para ello, nos apoyamos con el comando **sts** que gráfica el estimador de Kaplan-Meier:

Syntax

sts `graph [,by(varname)]`

donde `varname` es una variable categórica que indica una característica de las unidades de análisis (por ejemplo, tamaño empresarial para el análisis de supervivencia de empresas).

Implementación en Stata

Para la estimación de modelos paramétricos se deb utilizar el comando **streg**.

Syntax

streg [indepvars] [if] [in] [, options]

donde indepvars se refiere a las covariantes. Tener en cuenta que aquí no se especifica la variable dependiente puesto que al inicio del análisis se seteo la data (setset). Las opciones más importantes son:

- distribution(): especifica la distribución ha ser utilizada para el análisis de supervivencia (exponential, gompertz, loglogistic, weibul, ...).
- time(): Especifica que el modelo sea ajustado en el contexto del modelo de riesgo acelerado en lugar del contexto del logaritmo del hazard.

Asimismo, para estimar el modelo de COX debemos utilizar el comando **stcox**.

Syntax

stcox [indepvars] [if] [in] [, options]