Análisis de Supervivencia con Stata 6.0

(Análisis histórico de acontecimientos)

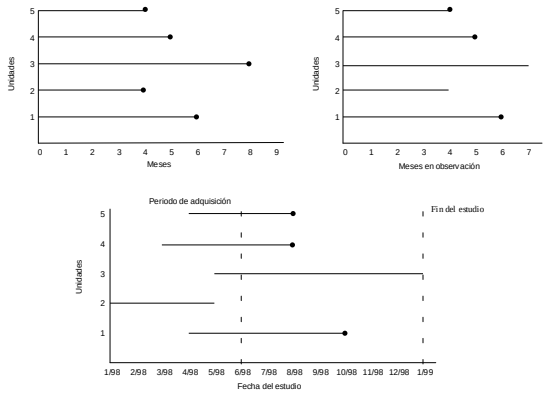
Presentado por
Mario Alberto Cleves Saa
Stata Corporation

Temas

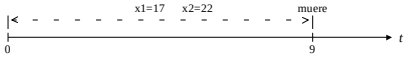
- **Conceptos básicos**
- **Definiciones importantes en Stata**
- **Funciones de supervivencia**
- **Ejemplos**
- **Análisis descriptivo y Métodos no paramétricos**
- **Modelo de los riesgos proporcionales de Cox**
- **Métodos paramétricos**
- **La orden stset**

Studios de Supervivencia

- Investigaciones en las cuales una muestra de unidades es observada por un período de tiempo durante el cual se producen uno o varios acontecimientos (en inglés, acontecimientos=events).
- Un estudio ideal comienza con la totalidad de la muestra y permanece con ella hasta que todas las unidades alcancen el objetivo determinado (el acontecimiento de interés).
- Sin embargo, la mayoría de los estudios no son ideales. Dos importantes excepciones caracterizan estos estudios:
  1. Hay investigaciones que necesitan muchas unidades o que investigan acontecimientos raros. Estos deben ir añadiendo unidades durante meses o años.
  2. Hay estudios en que unidades que se pierden o desaparecen del estudio o que no han sufrido el acontecimiento de interés antes de que el estudio termine. Estas observaciones se dicen que son censuradas.

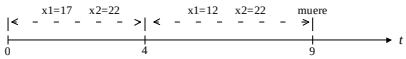


- Los datos de supervivencia son un conjunto de observaciones, cada una representando un episodio de tiempo, al final del cual ocurre o no el acontecimiento estudiado.
- Asociadas con cada observación pueden haber otras variables (covariables).
- En el más sencillo de los casos, cada observación representa los datos de un individuo.



id	tiempo	x1	x2	muerte	
101	9	17	22	1	(0,9]

- En casos más complicados, más de una observación le pertenece al mismo individuo



id	t0	tiempo	x1	x2	muerte	
101	0	4	17	22	0	(0,4]
101	4	9	12	22	1	(4,9]

- Al primer caso lo llamamos "datos con observaciones sencillas" (single-record data) y al segundo "datos con observaciones múltiples" (multiple-record data).

Definiciones importantes en Stata

- **Acontecimiento** (event) - Ocurren en un instante de tiempo.
- **Acontecimiento de interés** (failure event) - El acontecimiento que se desea analizar.
- **En riesgo** (at risk) - El individuo está en riesgo de que el acontecimiento de interés ocurra.
- **Origen** (Origin) - Momento en el cual el individuo entra en riesgo por primera vez..
- En Stata existen dos conceptos de tiempo:
  1. Tiempo - tal como es medido en el estudio. (días, fechas, etc.)
  2. Tiempo analítico (analysis time)

  Estos están relacionados por medio de la ecuación:

$$\textbf{Tiempo analítico} = \frac{\textbf{Tiempo - Origen}}{\textbf{Escala}}$$

- **Escala** (scale) - Un número fijo usado para convertir el tiempo a tiempo analítico
- La orden **stset** verifica los datos y crea el valor del tiempo analítico. Este nuevo valor es guardado en la variable _t.
- Simplemente, el tiempo analítico es el tiempo que transcurre desde que el individuo entra en riesgo.

- **Tiempo en observación** (under observation) - Tiempo durante el cual el individuo esta inscrito en el estudio.
- **Momento de entrada** (entry time) - momento en el que el individuo entra por primera vez al periodo en observación. Esto puede ocurrir antes, después o al mismo tiempo de que el individuo entre en riesgo.
- **Momento de salida** (exit time) - momento en el que el individuo sale por completo del periodo en observación.
- **Entrada retrasada** (delayed entry) - El individuo entra en observación después de entrar en riesgo.
- **Tiempo0** (time0) - Solo se usa cuando hay varias observaciones para cada individuo.
- **Brecha** (gap) - Intervalo de tiempo en el cual el individuo no esta en riesgo.
- **Pasado** (past) - Información anterior a que el individuo este bajo riesgo o observación.
- **Futuro** (future) - Información posterior al individuo salir de riesgo.

Clases de censuras

- **Censurado por la derecha** (Right censored) -El tiempo en observación cesó antes de que el acontecimiento de interés ocurriera.
- **Censurado por la izquierda (**left censored**)**- El evento ocurrió antes de que el individuo fuera observado.
- **Censurado en intervalos (interval censored**)- El evento ocurrió durante un intervalo de tiempo, pero no se sabe cuando exactamente.

- Con Stata se pueden analizar bases de datos con las siguientes características:

- Variables que varían en el tiempo
- Entradas retrasadas (left truncation)
- Escalas de tiempo múltiples
- Acontecimientos múltiples por unidad o sujeto
- Brechas (gaps)
- Observaciones censuradas por la derecha

Funciones de supervivencia

- Sea **T** una variable aleatoria que representa la duración (el tiempo de supervivencia).
- **Función de supervivencia (**survivor function**):** $\mathbf{S(}t\mathbf{)} = \mathbf{Pr(T} > t\mathbf{)}$
- Función de repartición **F(t)** de **T:** $\mathbf{F(}t\mathbf{)} = \mathbf{Pr(T} \leq t\mathbf{)} = \mathbf{1 - S(}t\mathbf{)}$
- Función de densidad es entonces: $f(t) = \frac{\mathbf{dF(}t\mathbf{)}}{\mathbf{d}t} = \frac{\mathbf{d}}{\mathbf{d}t}\mathbf{1 - S(}t\mathbf{)} = -\mathbf{S}'(t)$
- **Función de riesgo** (hazard function) define el riesgo instantáneo del acontecimiento (conditional hazard rate):

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbf{Pr}(t + \Delta t > T > t \mid T > t)}{\Delta t}$$

- Función de riesgo acumulado es la integral de la función de riesgo

$$H(t) = \int_0^t h(x)dx$$

Relación entre las funciones de supervivencia

- Todas las funciones de supervivencia están relacionadas entre sí :

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

$$f(t) = S(t)$$

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

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

Ejemplo 1 - todos mueren

- Estamos investigando una nueva medicina que promete alargar la vida de mujeres con cáncer del cuello uterino.
- Primero vamos a hacer un pequeño experimento aleatorio en el cual se suministra la nueva medicina a 25 ratas y un placebo a otras 25 ratas.
- Todas las ratas reciben la medicina al mismo tiempo y todas son seguidas hasta la muerte. No hay observaciones censuradas.

rata	medicina	tiempo	murió
1	placebo	11	1
2	placebo	11	1
3	nueva	11	1
4	placebo	12	1
5	placebo	12	1
	...		

- La orden **stset** para este caso es simplemente:

```
. stset  tiempo
```

Ejemplo 1 - stset

```
. stset tiempo

     failure event:  (assumed to fail at time=tiempo)
obs. time interval:  (0, tiempo]
 exit on or before:  failure
------------------------------------------------------------------------------
       50  total obs.
        0  exclusions
------------------------------------------------------------------------------
       50  obs. remaining, representing
       50  failures in single record/single failure data
      802  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        39
```

Errores producidos por stset

- Cuando los datos son **stset**, se ejecutan varias pruebas para verificar los datos. **stset** nos informa cuales observaciones tienen problemas y las marca con **_st**=0. Solo observaciones con **_st**=1 son incluidas en los análisis.

```
ignored because patid missing
event time missing                                PROBABLE ERROR
entry time missing                                PROBABLE ERROR
entry on or after exit (etime>t)                  PROBABLE ERROR
obs. end on or before enter()
obs. end on or before origin()
multiple records at same instant (t[_n-1]==t)     PROBABLE ERROR
overlapping records (t[_n-1]>entry time)          PROBABLE ERROR
weights invalid                                   PROBABLE ERROR
event time missing                                PROBABLE ERROR
entry time missing                                PROBABLE ERROR
```

Ejemplo 2 - observaciones censuradas

- En una investigación comparando dos tratamientos para el cáncer de la vejiga, 38 pacientes recibieron una medicina experimental y 47 pacientes recibieron placebo.
- Antes de entrar en el estudio, cada paciente tuvo cirugía con el fin de remover todos los tumores superficiales. El acontecimiento de interés es la primera reaparición del tumor.

Paciente	meses	tratamie	número	tamaño	r1
1	10	1	5	1	0
2	6	2	4	1	1
3	14	1	1	1	0
4	18	2	1	1	0
5	5	1	1	3	1
6	12	1	1	1	1
7	23	1	3	3	0

- La orden **stset** para este caso es también sencilla:

```
. stset  meses, fail(r1)
```

Ejemplo 2 - stset

```
. stset   meses, fail(r1)

     failure event:  r1 ~= 0 & r1 ~= .
obs. time interval:  (0, meses]
 exit on or before:  failure
------------------------------------------------------------------------------
       85  total obs.
        0  exclusions
------------------------------------------------------------------------------
       85  obs. remaining, representing
       47  failures in single record/single failure data
     1555  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        59
```

Ejemplo 3 - brechas

- En esta investigación vamos a evaluar el beneficio aportado por un nuevo protector experimental, diseñado para reducir la incidencia de fracturas de cadera en mujeres mayores de 60 años.
- Suponemos que durante episodios de tiempo en los cuales la mujer está hospitalizada, ella no está en riesgo del acontecimiento.

paciente	t0	meses	fractura	protecc	edad	calcio
16	0	5	0	1	77	7.78
16	5	12	1	1	77	9.73
17	0	8	0	1	66	11.48
17	8	15	1	1	66	10.79
18	0	5	0	1	64	11.58
18	15	17	1	1	64	11.59

Ejemplo 3 - stset

```
. stset  meses, fail(fractura) id(paciente) time0(t0)

               id:  paciente
     failure event:  fractura ~= 0 & fractura ~= .
obs. time interval:  (t0, meses]
 exit on or before:  failure

------------------------------------------------------------------------------
      106  total obs.
        0  exclusions
------------------------------------------------------------------------------
      106  obs. remaining, representing
       48  subjects
       31  failures in single failure-per-subject data
      714  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        39
```

Ejemplo 4 - Origen, fecha de entrada y fecha de salida

- Estamos interesados en investigar la relación entre la cantidad promedia de calorías consumidas por día y la incidencia de cardiopatía isquémica (enfermedad coronaria).

id	altenerg	evento	estatura	peso	fde	fds	fdn	ejer
14	0	0	172.01	89.09	16Dec2059	01Dec2076	03Jan2016	1
15	1	0	163.83	59.65	16May2062	20Aug2076	21Aug2006	1
16	0	3	171.20	89.40	16May2059	31Dec2059	16Sep1996	1
17	1	12	176.53	85.73	16Feb2059	14Jan2065	07May1999	0
18	1	0	177.80	94.80	16Feb2059	08Mar2068	09Mar1998	1

- La muestra tiene 337 sujetos. La variable **altenerg** indica si la persona consume un promedio mayor a 2.750 kcals/día.
- Las variables **fde**, **fds** y **fdn** son las fechas de entrada y salida del estudio, y la fecha de nacimiento respectivamente.
- La variable **evento** igual a 1, 3 o 13 indica cardiopatía isquémica. Otros valores indican otras enfermedades como cáncer y 0 indica que no hubo ningún acontecimiento.

Ejemplo 4 - stset

- Hay varias maneras de medir el tiempo en esta investigación. Podemos medir el tiempo desde el momento en el que la persona entrar al estudio o podemos medir el tiempo desde la fecha de nacimiento.
- En el primer caso tenemos:

```
. stset fds, fail(evento==1 3 13) origin(time fde) scale(365.25)

     failure event:  evento == 1 3 13
obs. time interval:  (origin, fds]
 exit on or before:  failure
    t for analysis:  (time-origin)/365.25
            origin:  time fde

--------------------------------------------------------------------------------
      337  total obs.
        0  exclusions
--------------------------------------------------------------------------------
      337  obs. remaining, representing
       46  failures in single record/single failure data
 4603.669  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  20.04107
```

Ejemplo 4 - stset

- En el segundo caso tenemos (fecha de nacimiento):

```
. stset fds, fail( evento==1 3 13) origin(fdn) scale(365.25) enter(time fde)

     failure event:  evento == 1 3 13
obs. time interval:  (origin, fds]
 enter on or after:  time fde
 exit on or before:  failure
    t for analysis:  (time-origin)/365.25
            origin:  time fdn

--------------------------------------------------------------------------------
      337  total obs.
        0  exclusions
--------------------------------------------------------------------------------
      337  obs. remaining, representing
       46  failures in single record/single failure data
 4603.669  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =  30.07529
                                  last observed exit t =  69.99863
```

Análisis descriptivo -stdes

- Stdes - describe brevemente los datos. No es una descripción analítica.
- Usando el tercer ejemplo (fracturas de cadera)

```
. stdes

                                |-------------- per subject --------------|
Category                  total        mean         min      median         max
--------------------------------------------------------------------------------
no. of subjects              48
no. of records              106    2.208333           1           2           3

(first) entry time                        0           0           0           0
(final) exit time                      15.5           1        12.5          39

subjects with gap             3
time on gap if gap           30          10           5          10          15
time at risk                714      14.875           1        11.5          39

failures                     31    .6458333           0           1           1
--------------------------------------------------------------------------------
```

Análisis descriptivo -stsum

- **stsum** - resume los datos de supervivencia.
- Usando el tercer ejemplo (fracturas de cadera)

```
. stsum

         |              incidence       no. of    |------ Survival time -----|
         | time at risk     rate        subjects        25%       50%       75%
---------+--------------------------------------------------------------------
   total |          714   .0434174            48          8        16        28


. stsum, by( proctecc)

         |              incidence       no. of    |------ Survival time -----|
proctecc | time at risk     rate        subjects        25%       50%       75%
---------+--------------------------------------------------------------------
       0 |          170   .1117647            20          4         8        12
       1 |          544   .0220588            28         22        28         .
---------+--------------------------------------------------------------------
   total |          714   .0434174            48          8        16        28
```

Análisis descriptivo - stvary

- **stvary** - Se usa para datos con observaciones múltiples. Demuestra cuales variables cambian en el tiempo.
- Usando el tercer ejemplo (fracturas de cadera)

```
. stvary

          subjects for whom the variable is
                                                      never    always   sometimes
variable |  constant    varying              missing   missing   missing
---------+----------------------------------------------------------------------
proctecc |        48          0                   48         0         0
    edad |        48          0                   48         0         0
  calcio |         8         40                   48         0         0
```

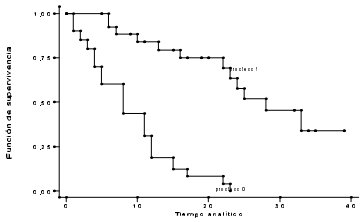
Análisis no paramétrico

- **sts** es un conjunto de órdenes que proporcionan métodos no paramétricos para el análisis de datos de supervivencia.
- Estas órdenes sirven para producir los valores de la función de supervivencia estimada usando el proceso del producto-límite de Kaplan-Meier y la función de riesgo acumulado de Nelson-Aalen. Así como otros valores relacionados incluyendo intervalos de confianza.
- Los valores producidos pueden incluirse en la base de datos como variables nuevas o pueden ser presentados como gráficas o listas.
- **sts** también incluye ordenes para comparar dos o más funciones de supervivencia.

Función de supervivencia

- Empezamos el análisis del tercer ejemplo estimando y dibujando las funciones Kaplan-Meier de supervivencia de los dos grupos experimentales.

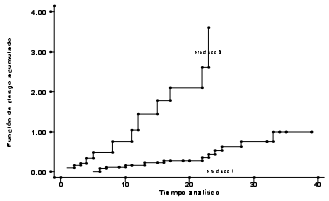
```
sts graph,by(proctecc)noborder t1(" ") l2("Función de supervivencia") b2("Tiempo
    analítico")
```



Función de riesgo acumulado

- Ahora usamos la opción **na** para dibujar la función Nelson-Aalen del riesgo integrado.

```
.sts graph, na by(proctecc)noborder t1(" ") l2("Función de riesgo acumulado") b2("Tiempo
    analítico")
```



Comparando funciones de supervivencia

- Las dos gráficas anteriores indican que el aparato experimental protege contra la fractura de la cadera. Sin embrago, deseamos probar esto estadísticamente.
- Con la orden **sts test** se pueden comparar dos o más funciones de supervivencia.
- **sts test** puede realizar la prueba del log-rank y de Wilcoxon (Breslow).
- La prueba del log-rank es la prueba predeterminada (default).

```
. sts test proctecc

Log-rank test for equality of survivor functions
------------------------------------------------
          |  Events
proctecc |  observed       expected
---------+------------------------
0        |        19           7.14
1        |        12          23.86
---------+------------------------
Total    |        31          31.00

               chi2(1) =      29.17
               Pr>chi2 =     0.0000
```

- La opción wilcoxon produce la prueba generalizada de Wilcoxon propuesta por Breslow.

```
. sts test  proctecc, wilcoxon

Wilcoxon (Breslow) test for equality of survivor functions
----------------------------------------------------------
          |  Events                        Sum of
proctecc |  observed       expected        ranks
---------+-------------------------------------
0        |        19           7.14          374
1        |        12          23.86         -374
---------+-------------------------------------
Total    |        31          31.00            0

               chi2(1) =      23.08
               Pr>chi2 =     0.0000
```

- Ambas son pruebas de rangos. Una importante diferencia entre estas prueba es que el log-rank le da un peso igual a todas los momentos donde ocurre el acontecimiento de interés, mientras que el Wilcoxon pesa más aquellos momentos con más observaciones.

Modelo de los riesgos proporcionales de Cox

- En el modelo Cox se asume que el riesgo del sujeto *i* es

$$\lambda_i(t) = \lambda_0(t)e^{\beta' X_i(t)}$$

- Donde λ_0 es el riesgo básico cuya forma funcional no es especificada. $\boldsymbol{X}_i$ es el vector de covariables del sujeto *i* (algunas que pueden variar con el tiempo) y $\boldsymbol{\beta}$ es un vector de coeficientes que debe ser estimado usando el método de verosimilitud parcial (partial likelihood).
- El modelo de Cox asume que el riesgo de dos sujetos con vectores de covariables fijas en el tiempo $\boldsymbol{X}_i$, y $\boldsymbol{X}_j$ es proporcional. Es decir

$$\mathbf{hr} = \frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)e^{\beta' X_i}}{\lambda_0(t)e^{\beta' X_j}} = \frac{e^{\beta' X_i}}{e^{\beta' X_j}}$$

- es constante en el tiempo.

Acontecimientos al mismo momento

- La verosimilitud parcial del modelo de riesgos proporcionales fue desarrollada asumiendo que el tiempo es continuo y por lo tanto dos sujetos no pueden sufrir el acontecimiento de interés al mismo tiempo.
- En realidad, como anotamos el tiempo de una manera discreta, es posible que para dos o más personas el acontecimiento de interés ocurra en el mismo momento.
- Cuando ocurren empates en el tiempo la verosimilitud parcial debe ser modificada. En Stata hay cuatro maneras de hacer esto. Las opciones son **breslow**, **efron**, **exactm** y **exactp**.
- Si no hay tiempos empatados en los datos, todos los métodos producen los mismos resultados. Si hay pocos empates, los resultados de los cuatro métodos son muy parecidos.
- El problema simplemente es saber la manera correcta de ordenar los acontecimientos que ocurren al mismo tiempo. Más específicamente se desea saber ¿quien esta en riesgo cada vez que alguien sufra el acontecimiento de interés?

- **Método de Breslow** - (**breslow**). Esta es la forma predeterminada de la orden **stcox**. En este método se usa el mayor número de casos en riesgo para calcular cada uno de los acontecimientos empatados. Produce los resultados rápidamente pero puede ser la menos exacta cuando hay muchos tiempos empatados.
- **Método de Efron** - (**efron**). Este método es preferido al de Breslow cuando hay muchos empates. El método puede ser mas lento que el de Breslow especialmente cuando se usa la opción **robust**.
- **Método de la verosimilitud parcial exacta** - (**exactp**). En este método hay que enumerar todo los conjuntos de riesgo posibles en cada momento que ocurre el acontecimiento de interés. Es por lo tanto el método más lento y puede demorarse mucho si hay muchos acontecimientos que ocurren al mismo tiempo.
- **Método de la verosimilitud marginal exacta**- (**exactm**). Este método produce valores similares a los que se obtienen con el Método de Efron. Este método también requiere que se haga enumeración completa, pero esta puede ser reemplazada por una evaluación numérica de una integral.

Ejemplo: Cirrosis biliar primaria (CBP)

- La CBP es una enfermedad crónica fatal del hígado que afecta principalmente a las mujeres.
- En este estudio investigamos los factores importantemente asociados con el riesgo de morir.

```
. stset días, failure(murió) scale(365.25)

     failure event:  murió ~= 0 & murió ~= .
obs. time interval:  (0, días]
 exit on or before:  failure
    t for analysis:  time/365.25
--------------------------------------------------------------------------
      312  total obs.
        0  exclusions
--------------------------------------------------------------------------
      312  obs. remaining, representing
      125  failures in single record/single failure data
 1713.854  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  12.47365
```

- **Variables**:

edad
sexo: 0=masculino, 1=femenino
presencia de **ascitis** 0=no 1=si
presencia de **hepatomegalia** 0=no 1=si
presencia de **edema** 0=no 1=si
bilirubina en mg/dl
colesterol en mg/dl
albúmina en mg/dl
cobre en la orina in ug/day
fosfatasa alcalina in U/liter
SGOT in U/ml
tiempo de **protrombina** en segundos
estado histológico de la enfermedad

- Al analizar estos datos se descubrió que las covariables importantes con relación a la muerte eran la edad, la presencia de edema, la bilirubina, el tiempo de la protrombina y la albúmina.

```
. stcox  edad edema bili prot albúmina, efron nolog noshow
Cox regression -- Efron method for ties

No. of subjects =          312                     Number of obs   =       312
No. of failures =          125
Time at risk    =  1713.853525
                                                   LR chi2(5)      =    159.17
Log likelihood  =   -560.37903                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |
          _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   1.033381   .0095645      3.548   0.000       1.014804    1.052298
       edema |   1.416384   .3210674      1.536   0.125       .9083008    2.208677
        bili |   1.126286   .0168733      7.938   0.000       1.093695    1.159847
        prot |   1.322025   .0945815      3.902   0.000       1.149059    1.521027
    albúmina |   .2716974   .0618736     -5.722   0.000       .1738768    .4245506
------------------------------------------------------------------------------
```

Modelos estratificados

- En Stata podemos estimar modelos estratificados. Esto es útil cuando el estudio esta compuesto de varios grupos que representan diferentes poblaciones.
- Por ejemplo, un estudio comparando dos medicinas para el tratamiento de una enfermedad es hecho simultáneamente en varios centros de investigación. Se creé que hay diferentes tipos de pacientes en cada centro, pero que el impacto de la medicina es proporcionalmente el mismo en todos los centros. Entonces el riesgo para el paciente i en el centro k es

$$\lambda_{ki}(t) = \lambda_{0k}(t)e^{\beta' X_i(t)}$$

- Se deja variar el riesgo básico de un grupo a otro, pero se obliga que los coeficientes sean iguales. En general:

$$\lambda_i(t) = \begin{cases} \lambda_{01}(t)e^{\beta' X_i(t)} & \text{sujeto en el primer strato} \\ \lambda_{02}(t)e^{\beta' X_i(t)} & \text{sujeto en el segundo strato} \\ \lambda_{03}(t)e^{\beta' X_i(t)} & \text{sujeto en el tercer strato} \end{cases}$$

Residuos

- En Stata se pueden obtener residuos eficientes de score (derivada), residuos martingales, residuos de Schoenfeld, residuos escalados de Schoenfeld, residuos de Cox-Snell y residuos de deviance.
- Estos sirven para examinar varios aspectos del modelo, incluyendo:
  la forma funcional de las covariables (martingales)
  la validación del modelo (Cox-Snell, martingales)
  el examen de puntos de influencia y outlier (Schoenfeld, score)
  la validación de la suposición de los riesgos proporcionales (Schoenfeld, score)
- Ejemplos: forma funcional de las covariables y del la suposición de los riesgos proporcionales.

Residuos martingales

- El residuo martingale de un sujeto es simplemente:

  (# de acontecimientos observados) - (# de acontecimientos esperados dado el modelo)
- El desarrollo y las propiedades de estos residuos son basados en la teoría de procesos contables (Fleming y Harrington, 1991, Counting Processes and Survival Analysis).
- Los residuos martingales de la regresión de Cox se obtienen vía la opción **mgale( ).**

```
. stcox edad edema bili prot albúmina, efron mgale(marting )
```

- La opción **mgale(marting)** crea la variable **marting** con el valor del residuo para cada observación**.**
- Si hay varias observaciones por sujeto el valor total del martingale para ese sujeto es la suma de los martingales parciales.

Forma funcional de las covariables
método sencillo

- El modelo especifica que

$$\lambda_i(t) = \lambda_0(t)e^{\beta' X_i(t)}$$

- ¿Es la forma de exp(β'X) correcta? Quizá debemos usar: log(X), X^2, o $I_{X>65}$
- Supóngase que el verdadero modelo es

$$\lambda_i(t) = \lambda_0(t)e^{f(x)}$$

- Si ***M*** sea el valor del residuo cuando no se incluyen covariables en el modelo de Cox, entonces se puede demostrar que

$$E(M_i) \approx \mathbf{c}\, f(x_i)$$

donde **c** es una constante que depende del número de observaciones censuradas.

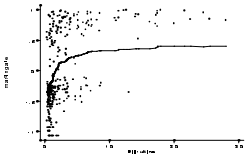
- Si hacemos una gráfica de ***x*** contra ***M*** podremos observar la forma de $f(x_i)$ ya que **c** solo afecta la escala del eje vertical (ordenada).

- Primero calculamos los residuos martingales de la regresión de Cox sin covariables.
- Usamos la opción **estimate.**

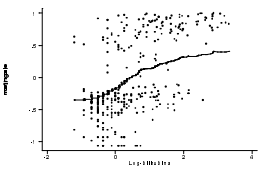
```
. stcox , efron nolog noshow mgale(marting) estimate

Cox regression -- Efron method for ties

No. of subjects =          312                     Number of obs   =       312
No. of failures =          125
Time at risk    =  1713.853525
                                                   LR chi2(0)      =      0.00
Log likelihood  =   -639.96649                     Prob > chi2     =         .
------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
------------------------------------------------------------------------------
```

- Ahora hacemos una gráfica de la variable **bilirubina** contra el residuo sobreponiendo un "smoother".

```
. ksm marting bili, ylab(-1, -.5, 0, .5, 1) xlab t1(" ") b2("Bilirubina")
```



```
. ksm marting lbili, ylab(-1, -.5,0,.5,1) xlab t1(" ") b2("Log-bilirubina")
```



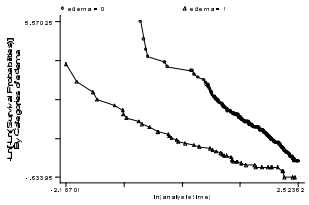
```
. stcox  edad edema  lalb lbili lpt, efron nolog noshow

Cox regression -- Efron method for ties

No. of subjects =          312                     Number of obs   =       312
No. of failures =          125
Time at risk    =  1713.853525
                                                   LR chi2(5)      =    195.95
Log likelihood  =    -541.9904                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    edad |    1.03314   .0089122      3.779   0.000       1.015819    1.050756
   edema |   1.493402   .3306777      1.811   0.070       .9676064    2.304914
    lalb |   .0396114   .0280211     -4.564   0.000        .009901     .158475
   lbili |   2.446095   .2397966      9.124   0.000       2.018496    2.964276
     lpt |   25.41678   25.37109      3.241   0.001       3.592932    179.8009
------------------------------------------------------------------------------
```

La suposición de los riesgos proporcionales

- La suposición más importante de la regresión de Cox es que los riesgos son proporcionales en el tiempo. Esto debe ser verificado.
- Supóngase que un estudio sigue por 10 años un grupo de pacientes, algunos de los cuales reciben un tratamiento experimental. Si el riesgo de morir en el grupo sin tratamiento es 3 veces el riesgo en el otro grupo (HR=3.0), la suposición de los riesgos proporcionales implica que este HR es igual el primer años, el segundo años y en cualquier tiempo durante el estudio.
- Stata tiene tres programas para evaluar esta suposición. **stphplot** y **stcoxph** son métodos gráficos que pueden ser útiles en algunas ocasiones.
- **stphtest** es una prueba analítica de la suposición.

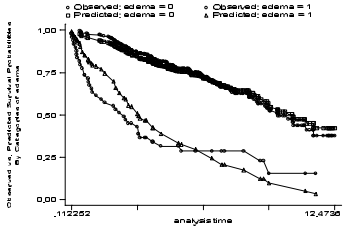
- La orden **stphplot** dibuja una curva del -ln(-ln(S(*t*))) contra ln(tiempo analítico) para cada valor de una variable nominal o ordinal. Si la suposición no es violada las curvas son más o menos paralelas.

```
. stphplot ,by(edema)
```



- **stcoxkm** produce una gráfica que compara la función de supervivencia obtenida con el método de Kaplan-Meier con la obtenida con Cox, para cada valor de una variable nominal o ordinal.

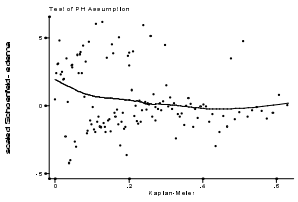
```
. stcoxkm ,by(edema) c(llll) s(OOST)
```



- La orden **stphtest** produce una prueba estadística de la suposición. Para obtener una prueba global del modelo es necesario que se hayan guardado los residuos de Schoenfeld al ejecutar **stcox**. Para obtener una prueba para cada covariable es necesario haber guardado los residuos *escalados* de Schoenfeld.

```
. stcox edad edema lalb lbili lpt, efron schoenfeld(sch*) scaledsch(sca*)
. stphtest, detail km

        Test of proportional hazards assumption

        Time:  Kaplan-Meier
        ------------------------------------------------------------------
                    |       rho           chi2       df       Prob>chi2
        ------------+-----------------------------------------------------
        edad        |    -0.04078         0.18        1          0.6721
        edema       |    -0.20243         4.54        1          0.0332
        lalb        |    -0.00786         0.01        1          0.9288
        lbili       |     0.15785         2.80        1          0.0941
        lpt         |    -0.20503         4.06        1          0.0439
        ------------+-----------------------------------------------------
        global test |                     12.35        5          0.0303
        ------------------------------------------------------------------
```

```
. stphtest, km plot(edema) lowess
```



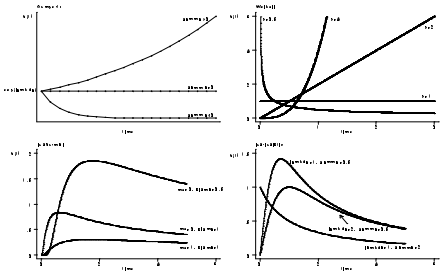
Métodos paramétricos

- En Stata se pueden estimar 6 modelos paramétricos:
- Exponencial, Weibull, Gompertz, log-normal, log-logístico y el gamma generalizado.

Distribución	Métrica	Función de Riesgo
Exponencial	RP, TA	constante
Weibull	RP, TA	monótona
Gompertz	RP	monótona
Log-normal	TA	variable
Log-logística	TA	variable
Gamma	TA	variable

RP = riesgo proporcional, TA = modelos de tiempo accelerado (accelerated failure time).

- Stata puede producir para los modelos paramétricos los mismo residuos que produce para Cox, con excepción de los residuos de Schoenfeld.



- Los seis modelos paramétricos se estiman usando la orden **streg**. Por ejemplo para estimar un modelo de Weibull

```
. streg  edad edema  lalb lbili lpt, dist(weibull) nolog

Weibull regression -- log relative-hazard form

No. of subjects =          312                     Number of obs   =       312
No. of failures =          125
Time at risk    =  1713.853525
                                                   LR chi2(5)      =    197.27
Log likelihood  =   -225.71794                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |    1.03246   .0087634      3.764   0.000       1.015427     1.04978
       edema |    1.56163   .3389512      2.054   0.040       1.020529     2.389633
        lalb |    .043355   .0300057     -4.535   0.000       .0111668     .1683257
       lbili |   2.364123   .2207607      9.214   0.000       1.968726     2.838932
         lpt |   32.89736   31.85967      3.607   0.000       4.929574     219.5395
-------------+----------------------------------------------------------------
       /ln_p |   .4550597   .0711444      6.396   0.000       .3156192     .5945003
-------------+----------------------------------------------------------------
           p |   1.576268   .1121427                          1.371108     1.812125
         1/p |   .6344101   .0451347                          .5518383     .7293371
------------------------------------------------------------------------------
```

- La función de riesgo gamma es muy flexible. Ella pueden acomodar muchas formas incluyendo como casos especiales el modelo de Weibull cuando κ=1, el exponencial cuando κ=1 y σ=1, y el modelo log-normal cuando κ=0. Este modelo es usado principalmente para evaluar y ayudar a selecionar el modelo paramétrico más adecuado.

```
streg edad edema lalb lbili lpt, dist(gamma) nolog noshow

Gamma regression -- accelerated failure-time form

No. of subjects =          312                     Number of obs   =       312
No. of failures =          125
Time at risk    =  1713.853525
                                                   LR chi2(5)      =    201.46
Log likelihood  =   -222.99397                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   -.0225886   .0060394     -3.740   0.000      -.0344256   -.0107516
       edema |   -.3740681   .1630701     -2.294   0.022      -.6936795   -.0544566
        lalb |    1.992172   .4838759      4.117   0.000       1.043793    2.940551
       lbili |   -.5592369   .0646372     -8.652   0.000      -.6859235   -.4325502
         lpt |   -2.638961   .7070007     -3.733   0.000      -4.024657   -1.253265
       _cons |    7.612197   1.821831      4.178   0.000       4.041474    11.18292
-------------+----------------------------------------------------------------
      /ln_sig |   -.2744318   .0930824     -2.948   0.003       -.45687   -.0919936
       /kappa |    .4945175    .198959      2.486   0.013        .104565     .88447
-------------+----------------------------------------------------------------
        sigma |    .7600038    .070743                          .6332627     .912111
------------------------------------------------------------------------------
```

- La prueba de la hipótesis nula κ=0 (que el modelo es log-normal) usando la prueba de Wald, es la que se presenta en los resultados, p=**0.013**. Esto sugiere que el modelo log-normal no es adecuado para estos datos.

- La prueba de chi-cuadrado de la hipótesis nula κ=1 se puede hacer fácilmente.

```
.testnl [kappa]_cons =1

  (1)  [kappa]_cons =1

               chi2(1) =        6.45
           Prob > chi2 =        0.0111
```

rechazando el modelo de Weibull.

- Estimemos los otros modelos paramétricos. Usamos la opción **time** para el modelo exponencial y el de Weibull para poder comparar los parámetros producidos por los diferentes modelos.

- Podemos compara los valores del logaritmo de la verosimilitud de los modelos usando el criterio de Akaike, el cual penaliza cada valor basándose en el número de parámetros estimados.

  Akaike = -2(log verosimilitud) + 2(c + p + 1)

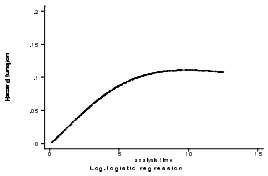
  donde c es el número de covariables y p el número de parámetros auxiliares.

	exponencial	Weibull	log-normal	log-logistico	gamma
edad	-0.028833	-0.020266	-0.022526	-0.024328	-0.022589
edema	-0.338735	-0.282776	-0.500870	-0.433900	-0.374068
log albúmina	2.309385	1.990999	1.783477	1.984961	1.992172
log bilirubina	-0.696257	-0.545851	-0.556026	-0.553956	-0.559237
log tp	-3.498407	-2.216243	-3.001654	-2.873695	-2.638961
constante	10.10519	6.596764	8.005586	8.104667	7.612197
parámetros		1.576268	0.888227	0.478490	0.494518
auxiliares					0.760004
Log verosimilitud	-242.44201	-225.71794	-226.30068	-222.24662	-222.99397
Akaike	496.88402	465.43588	466.60136	458.49324	461.98794

- Según el criterio de Akaike el modelo log-logístico es preferido ya que tiene el menor valor de Akaike.

```
. streg edad edema lalb lbili lpt, dist(loglog) nolog noshow

Log-logistic regression -- accelerated failure-time form

No. of subjects =          312                     Number of obs   =       312
No. of failures =          125
Time at risk    =  1713.853525
                                                   LR chi2(5)      =    207.98
Log likelihood  =   -222.24662                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   -.024328   .0062638     -3.884   0.000      -.0366048   -.0120512
       edema |  -.4339001   .1748166     -2.482   0.013      -.7765344   -.0912659
        lalb |   1.984961   .5069562      3.915   0.000       .9913451    2.978577
       lbili |  -.5539557   .0678524     -8.164   0.000      -.686944    -.4209675
         lpt |  -2.873695   .7535226     -3.814   0.000      -4.350573   -1.396818
       _cons |   8.104667   1.924106      4.212   0.000       4.333488    11.87585
-------------+----------------------------------------------------------------
      /ln_gam |  -.7371192   .0722599    -10.201   0.000      -.8787459   -.5954925
-------------+----------------------------------------------------------------
        gamma |   .4784904   .0345756                          .4153034     .551291
------------------------------------------------------------------------------
```

```
. stcurv , haz at(edema=0) c(l) s(.)
```




```
. stcurv , haz at(edema=1) c(l) s(.)
```