Modelos de Mercadotecnia
Modelos de Supervivencia

Jorge de la Vega Góngora

Maestría de Mercadotecnia,
Instituto Tecnológico Autónomo de México

Sesión 9

Introducción

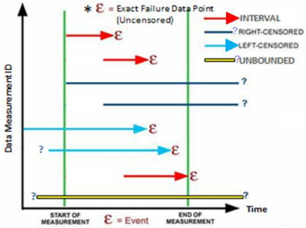
Introducción

- El modelado del **tiempo de ocurrencia de eventos** es un tópico importante que tiene muchas aplicaciones en áreas muy diversas.
- Si el tiempo hasta el que ocurre un evento no fuera importante, simplemente tendríamos que contar cuántos eventos han ocurrido en un intervalo de tiempo y utilizar una variable binaria para registrarlo. Pero en ciertas situaciones, cuando el evento es crítico, resulta importante tomar en cuenta el tiempo antes de su ocurrencia.
- La definición de evento puede ser muy variable e incluye muerte, graduación, compras, abandono (*churn*), bancarrota, incumplimiento, etc.
- Los métodos que se usan para analizar esos datos se conocen como **Análisis de Supervivencia**, y buscan responder a la pregunta: ¿Cuánto tiempo pasará antes de que un evento ocurra? El nombre viene de su aplicación origen que son las áreas de la salud, pero su uso se ha extendido de manera importante a otras áreas.
- Otros nombres que se pueden encontrar incluyen:
    - *Análisis de historia de eventos* (ciencias sociales)
    - *Análisis de duraciones* (econometría)
    - *Análisis de confiabilidad* (ingeniería).

- En el caso de mercadotecnia, el análisis del ciclo de vida de un cliente utiliza información como el tiempo entre compras, para predecir el valor potencial de un cliente. Algunas aplicaciones incluyen:
  - identificación de clientes potencialmente estancados (no regresan a comprar)
  - En un esquema de renovación o suscripción, un indicador clave es el tiempo que un cliente se mantiene en activo más allá de su fecha de expiración. (vida residual esperada).
  - La identificación de estos clientes ayuda a dirigir campañas adecuadas y específicas para mantener su interés.
- DuWORS y Haines (1990) Utilizan el análisis de supervivencia, en particular la función de riesgo, para medir la lealtad del consumidor a una marca y muestran que la lealtad es variable en el tiempo.
- Jin, et.al (2021) crean un modelo de predicción de el precio de reservación: es el precio más alto que está dispuesto a pagar un consumidor por una unidad de un servicio o producto específico. El modelo trata más con precios censurados más que con tiempos censurados, y puede incluir censura por la izquierda y por la derecha.

- Ganchev et.al (2012) Trata el problema conocido como *dark pool problem*. Los *dark pools* son mercados de intercambio en donde los traders buscan el intercambio "invisible" de grandes volúmenes de activos a precios de mercado. Otras aplicaciones se han llevado a cabo en el área de subastas.
- Ansell, Harrison y Archibald (2007) Usan segmentación de estilos de vida y análisis de supervivencia para identificar oportunidades de ventas cruzadas (las ventas cruzadas se refieren a la estrategia de vender otros productos a un cliente que ya ha comprado un producto del vendedor).

- El análisis de supervivencia estudia variables aleatorias positivas $T > 0$, que representan el tiempo que transcurre entre la ocurrencia de eventos.
- El objetivo del análisis de supervivencia incluye:
  1. Estimar e interpretar la función de supervivencia
  2. Comparar funciones de supervivencia de diferentes segmentos o grupos de sujetos
  3. Evaluar la relación entre la supervivencia y uno o más predictores.
- La principal diferencia entre las técnicas del Análisis de supervivencia y otras técnicas estadísticas, es la presencia de información parcial o incompleta, como las observaciones censuradas o truncadas.



Observaciones censuradas I

Información censurada

Se dice que una observación es **censurada** cuando no se conoce el tiempo T exacto de ocurrencia del evento y unicamente se sabe que ocurrió en uncierto intervalo. Los datos usualmente consisten de n observaciones independientes correspondientes con pares de puntos (t_i, δ_i), donde

- δ_i es una variable indicadora del status de ocurrencia del evento ($\delta_i = 0$ si el evento no se observó para el sujeto i, y $\delta_i = 1$ si el evento se observó).
- t_i es el tiempo al evento (si $\delta_i = 1$) o el tiempo censurado (si $\delta_i = 0$). A t_i también se le conoce como el tiempo de duración.

- Entonces los datos observados que se tienen son una muestra de pares de valores $(t_1, \delta_1), (t_2, \delta_2), \ldots, (t_n, \delta_n)$. Es común representar los datos como t_i para $(t_i, 1)$ o t_i^+ para $(t_i, 0)$.
- El que una observación sea censurada no tiene que ver con la calidad de los datos, sino con la propia naturaleza de éstos. Usualmente significa que la observación del experimento se termina antes de la ocurrencia del evento.
- Entre las razones de censura de los datos, se tiene principalmente:
  - El individuo se sale (o lo sacan) del estudio
  - El estudio tiene un tiempo fijo y el evento relevante no ocurre durante el periodo del estudio

Observaciones censuradas II

- Este tipo de censura se conoce como *censura por la derecha*. Los tipos de censura se definen a continuación.
  - **Censura por la derecha**: Solamente se observa un tiempo C, menor al tiempo de fallo exacto T. Hay varios tipos de cesurado por la derecha, pero no se considerarán aquí.
  - **Censura por la izquierda**: Solamente se sabe que el tiempo de fallo ocurrió antes de un tiempo observado C. Por ejemplo, cuando un sistema registra los precios por debajo de un límite como ceros.
  - **Censura por intervalo**: cuando solamente se sabe que una observación se encuentra dentro de un intervalo de tiempo pero no se conoce con precisión. Por ejemplo, cuando se hace revisiones periódicas de las bases de datos de clientes y algunos clientes tuvieron un evento entre una revisión y otra.

- **Análisis de retención de clientes**. En los esquemas de suscripción es un tema importante. Un objetivo es tratar de pronosticar las tasas de retención para poder predecir valor de los clientes. Supongamos que se tienen 42 clientes y se quiere evaluar el impacto de una promoción para mantener su lealtad. Cada cliente se aleatoriza en uno de dos grupos: los que reciben la promoción y los que no. El estudio se termina un año después. Los clientes que tienen diferentes tiempos porque se incorporaron en diferentes fechas en el programa. Los tiempos son en semanas:

promoción	$6, 6, 6, 7, 10, 13, 16, 22, 23, 6^+, 9^+, 10^+, 11^+ 17^+, 19^+, 20^+, 25^+, 32^+, 32^+ 34^+, 35^+$
no promoción	$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$

El evento que se considera es el de renovación de suscripción. Por ejemplo 17^+ significa un caso que se registro 17 semanas antes de la terminación del estudio y no ha renovado su suscripción.

- La función de supervivencia de una variable aleatoria $T > 0$ se define como

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

  donde $S(t)$ es la proporción de sujetos sin evento hasta t y $F(t)$ es la función de distribución usual de la variable aleatoria T.

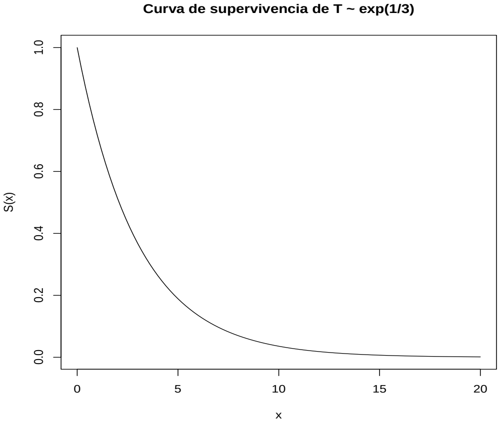
- Como ejemplo, consideremos $T \sim \exp(\lambda)$. En este caso, la función de densidad de T es $f(t) = \lambda e^{-\lambda t}$, su función de distribución es

$$F(t) = \int_0^t f(x)dx = 1 - e^{-\lambda t} = 1 - S(t)$$

  por lo que la función de supervivencia es $S(t) = e^{-\lambda t}$. En este caso $\mathrm{E}(T) = \frac{1}{\lambda}$. Si tomamos por ejemplo $\lambda = 1/3$.

```
curve(1-pexp(x, rate = 1/3), from = 0, to = 20, main = "Curva de supervivencia de T ~ exp(1/3)", ylab = "S(x)")
```

Función de supervivencia II



Ejemplo: precio de reservación I

- Se desea establecer el precio de un producto ω para alcanzar la máxima ganancia cuando se vende a una población
- Si se conoce el precio de reservación r_i de cada consumidor i, la **Función de probabilidad de compra** sobre el precio ν está dada por

  $$FPC(\nu) = \frac{1}{n}\sum_{i=1}^{n} I(r_i \geq \nu)$$

  donde n es el número total de consumidores.
- Con la función FCP, se puede alcanzar la ganancia máxima esperada buscando el precio del producto ω como

  $$\nu^* = \arg\max_{\nu}\{(v - c)FPC(\nu)\}$$

  donde c es el costo de producción de ω.
- Cuando R_i es el precio de reserva considerado como variable aleatoria, tenemos que $FPC(x) = P(R_\omega > \nu)$, por lo que la función de probabilidad de compra es una función de supervivencia.

Estimación no paramétrica (Kaplan-Meier) de $S(t)$

- El estimador de Kaplan-Meier es un estimador empírico no paramétrico. También se conoce como el método de estimación producto-límite.
- Usualmente se estima con un modelo no paramétrico, la estimación de Kaplan-Meier, que se define como:

$$\hat{S}_{KM}(t) = \prod_{i:t_i<t}\left(1 - \frac{d_i}{n_i}\right)$$

  donde:
  - n_i = Número de sujetos presentes al tiempo t_i (sujetos en riesgo)
  - d_i = número de eventos que ocurren en el tiempo t_i
- Por ejemplo, para los datos de promoción mostrados previamente:

t_i	n_i	d_i	$(1-d_i/n_i)$	$\hat{S}_{KM}(t_i)$
6	21	3	0.8571	0.8571
7	17	1	0.9412	0.8067
10	15	1	0.9333	0.7529
13	12	1	0.9167	0.6902
16	11	1	0.9091	0.6275
22	7	1	0.8571	0.5378
23	6	1	0.8333	0.4482

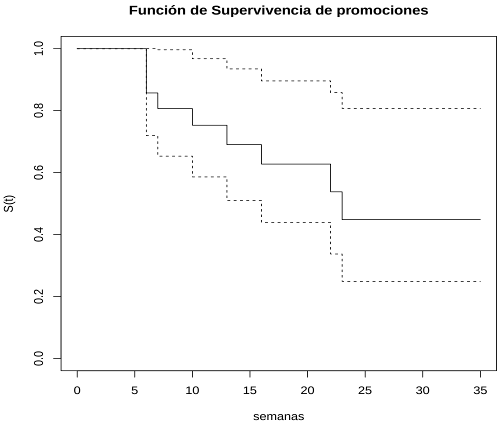
- Para hacer el ejercicio en R, necesitamos definir los tiempos con su respectiva variable dummy δ.
- La función Surv crea los pares de variables (t_i, δ_i) para usarlos como respuesta en el modelo correspondiente.

```
datos <- data.frame(prom = c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,32,34,35),
                    renov = c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0))
m1 <- survfit(Surv(prom,renov) ~ 1, data = datos)
summary(m1)
```

```
Call: survfit(formula = Surv(prom, renov) ~ 1, data = datos)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807
```

```
plot(m1, main = "Función de Supervivencia de promociones", xlab = "semanas", ylab = "S(t)")
```

Estimación no paramétrica II



- En este ejemplo, la tasa de no renovación de 6 semanas es 85.71 % y la tasa de no renovación de 22 semanas es 53.78 %. El cálculo de los intervalos de confianza se da con la siguiente fórmula, que se obtiene como una aproximación de la estimación por máxima verosimilitud de la función de supervivencia. El intervalo es de la forma:

$$\hat{S}(t)\,\mathrm{exp}(\pm 1.96\hat{s}(t))$$

donde

$$\hat{s}^2(t) = \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

- Por ejemplo, para calcular el error estándar de $\hat{S}(7)$ se calcula como:

$$se(\hat{S}(7)) = 0.8067\left[\frac{3}{(21)(18)} + \frac{1}{(17)(16)}\right]^{1/2} = 0.0869$$

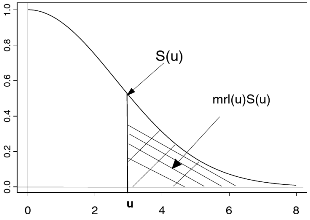
y el intervalo de confianza del 95 % para $S(7)$ es $(0.653, 0.996)$.

- La función de supervivencia es una probabilidad, por lo que tiene las siguientes propiedades
  - $t \in [0, \infty)$
  - $S(t)$ es no creciente: $S(t_1) \geq S(t_2)$ para $t_1 \leq t_2$
  - $S(0) = 1$
  - $S(t) = 1 - F(t)$ donde F es la función de distribución de T.
- Cuando $S(t)$ es continua, se puede escribir como $S(t) = \int_t^\infty f(\tau)d\tau$. En este caso se tiene además que $f(t) = -S'(t)$ y

$$E(T) = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$$

  Entonces el tiempo esperado es el área bajo la curva $S(t)$.
- En el ejemplo exponencial, $E(T) = 1/\lambda$, que es el área bajo la curva de la función de supervivencia.

- **Vida residual media**: Se define como $r(t) = E(T - t|t \leq T) = \frac{\int_t^\infty S(\tau)d\tau}{S(t)}$. A veces le llaman `mrl(t)`. Para los casos que tienen edad t, mide su vida restante promedio. Noten que $r(t)$ es el área bajo la curva a la derecha de t, dividida por $S(t)$.



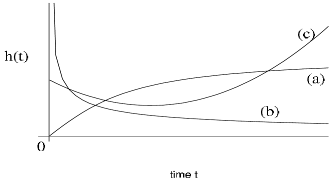
- **Función de riesgo o fuerza de mortalidad**: $\lambda(t)$ nos da la tasa de ocurrencia de eventos instantánea, suponiendo que los individuos no han tenido evento hasta t:

$$\lambda(t) = \lim_{\delta \to 0} \frac{P(t \leq T \leq t + \delta | t \leq T)}{\delta}$$

En otras palabras: es la probabilidad de que el evento se dé en los siguientes pocos segundos, dado que no se ha dado hasta el momento t.

Funciones de $S(t)$ II

- La función de riesgo aproxima la proporción de sujetos que tienen eventos por unidad de tiempo alrededor de t. Notar que:
    - es una probabilidad condicional
    - es una medida de la propensión al evento como función de la edad del sujeto.
    - La función de riesgo, nos ayuda a entender el mecanismo de ocurrencia del evento (falla, muerte, abandono, etc.)



- Podemos ver algunas relaciones entre las diferentes funciones.

Funciones de $S(t)$ III

- La función de riesgo se relaciona con las otras funciones que ya definimos del siguiente modo:

$$h(t) = \lambda(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -(\log S(t))'$$

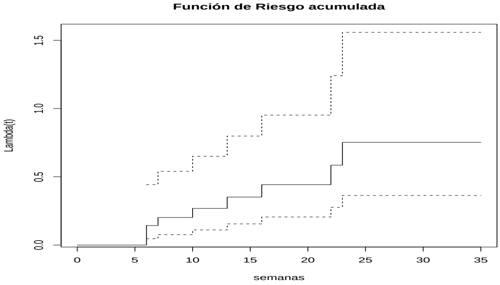
y

$$S(t) = \exp\left\{-\int_0^t \lambda(\tau)d\tau\right\} = \exp\{-\Lambda(t)\}$$

en donde la función $\Lambda(t) = \int_0^t \lambda(\tau)d\tau$ se le llama el **riesgo acumulado**.

- Podemos obtener una gráfica de la función de riesgo acumulado de la siguiente manera

```
plot(m1, cumhaz = T, xlab = "semanas", ylab = "Lambda(t)", main = "Función de Riesgo acumulada")
```



- No tenemos una función en `R` que calcule directamente la vida residual media pero podemos usar el archivo de funciones auxiliares para calcular las funciones de riesgo. La función `hazard.km` toma un objeto ajustado `survfit` y nos devuelve $\hat{\lambda}(t)$, $\hat{\Lambda}(t)$ y sus errores estándar, entre otros valores

```
source("../scripts/TK.R.functions.R.txt")  # Carga una lista de funciones complementarias
hazard.km(m1) # Calcula funciones de riesgo
```

```
  time ni di  hihat hitilde   Hhat se.Hhat Htilde se.Htilde
1    6 21  3 0.1429  0.1429 0.1542  0.0891 0.1429    0.0825
2    7 17  1 0.0196  0.0588 0.2148  0.1078 0.2017    0.1013
3   10 15  1 0.0222  0.0667 0.2838  0.1280 0.2683    0.1213
4   13 12  1 0.0278  0.0833 0.3708  0.1548 0.3517    0.1471
5   16 11  1 0.0152  0.0909 0.4661  0.1818 0.4426    0.1730
6   22  7  1 0.1429  0.1429 0.6202  0.2384 0.5854    0.2243
7   23  6  1     NA  0.1667 0.8026  0.3003 0.7521    0.2795
[1] "hazard.km:done"
```

- De acuerdo al modelo propuesto por DuWors y Haines, el modelo de supervivencia considerado hace las siguientes asociaciones:
  - El evento que se registra es el cambio de comportamiento en la compra de un consumidor. Es decir, el consumidor deja de comprar la marca. Se mide a través del registro en el estante en dos periodos.
  - La función de riesgo asociada es la medida natural de la lealtad de las personas. $\lambda(t)$ mide la tasa a la que consumidores que han comprado una marca en paricular, cambian a otra marca. Entonces es una medida directa pero inversa de la lealtad del cliente.

Ejemplos de modelos paramétricos: modelo exponencial

- Si se toma la función de riesgo $\lambda(t) = \lambda$ constante sobre el rango de T, la función de supervivencia es entonces $S(t) = e^{-\lambda t}$.
- Con este modelo, $\mathsf{E}(T) = 1/\lambda$. la función de riesgo $\lambda(t) = -(\log S(t))' = -(\log(\exp(-\lambda t)))' = \lambda$. Entonces la función de riesgo es constante igual a λ.
- El riesgo acumulado es $\Lambda(t) = \lambda t$.
- El parámetro del modelo se puede estimar con el siguiente comando:

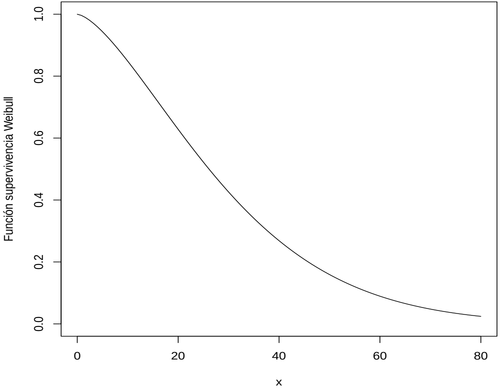
```
expo <- survreg(Surv(prom,renov)~1, data = datos, dist = "exponential")
summary(expo)
```

```
Call:
survreg(formula = Surv(prom, renov) ~ 1, data = datos, dist = "exponential")
            Value Std. Error    z      p
(Intercept) 3.686      0.333 11.1 <2e-16

Scale fixed at 1

Exponential distribution
Loglik(model)= -42.2   Loglik(intercept only)= -42.2
Number of Newton-Raphson Iterations: 4
n= 21
```

- Es una generalización de la distribución exponencial que permite una dependencia de una potencia del riesgo en el tiempo, y modelar de mejor manera la longitud de vida de humanos o animales.
- Se toma: $\lambda(t) = \lambda p(\lambda t)^{p-1}$ para $\lambda, p > 0$, La función de riesgo acumulada es $\Lambda(t) = (\lambda t)^p$ y la función de supervivencia es entonces $S(t) = \exp(-(t\lambda)^p)$.
- La distribución exponencial es un caso particular con valor $p = 1$. En general, si $T \sim \exp(\lambda)$, entonces $T^p \sim Weib(\lambda, p)$
- Con este modelo, $\mathrm{E}(T) = \frac{\Gamma(1+1/p)}{\lambda}$.
- En R las funciones `dweibull` y `pweibull` calculan las funciones de densidad y distribución respectivamente. Estas funciones usan los argumentos `shape` y `scale` para representar los parámetros p y $1/\lambda$ respectivamente. Por ejemplo, para $p = 1.5$ y $\lambda = 0.03$

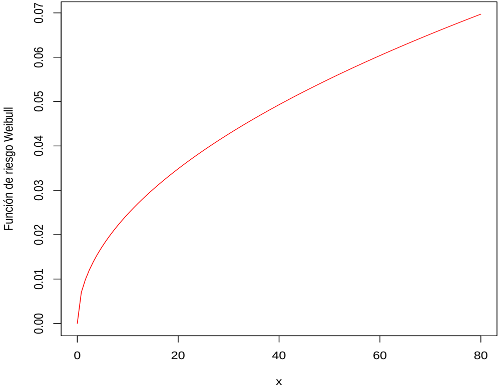
```
weibSurv <- function(t,shape,scale) pweibull(t,shape=shape, scale = scale, lower.tail = F) # definimos la función de supervivencia
curve(weibSurv(x, shape = 1.5, scale = 1/0.03), from = 0, to = 80, ylim = c(0,1), ylab = "Función supervivencia Weibull")
```

Ejemplos de modelos paramétricos: modelo Weibull II



- Para graficar la respectiva función de riesgo

```
weibhaz = function(t,shape,scale)dweibull(t,shape=shape, scale = scale)/pweibull(t,shape=shape,scale=scale, lower.tail = F)
curve(weibhaz(x, shape = 1.5, scale = 1/0.03), from = 0, to = 80, ylab = "Función de riesgo Weibull", col = "red")
```



Ejemplos de modelos paramétricos: modelo Weibull IV

- El modelo puede ser estimado en R del siguiente modo:

```
datos <- data.frame(prom = c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,32,34,35),
                    renov = c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0))
Weibull <- survreg(Surv(prom,renov)~1, data = datos, dist = "weibull")
summary(Weibull)
```

```
Call:
survreg(formula = Surv(prom, renov) ~ 1, data = datos, dist = "weibull")
             Value Std. Error     z      p
(Intercept)  3.519      0.273 12.87 <2e-16
Log(scale)  -0.303      0.278 -1.09   0.28

Scale= 0.739

Weibull distribution
Loglik(model)= -41.7   Loglik(intercept only)= -41.7
Number of Newton-Raphson Iterations: 5
n= 21
```

- El mapeo de parámetros que usa R es que define el parámetro de escala como $\sigma = 1/p$ y el parámetro de media como $\mu = -\log(\lambda)$

Ejemplo: Retención de clientes I

- El valor del tiempo de vida de un cliente (*customer lifetime value* o *LTV*) se define usualmente como el ingreso total neto que una compañía puede esperar de un cliente (Novo 2001).
- El *LTV* juega un papel relevante en varias aplicaciones, principalmente:
  - Análisis de abandono y retención: el LTV complementa la probabilidad de abandono, indicando cuánto se pierde por el abandono y cuánto esfuerzo se requiere poner en el segmento.
  - Campañas de retención: relación entre recursos invertidos y el correspondiente cambio en el LTV de los segmentos objetivo.
- A nivel secundario el *LTV* se puede utilizar en Análisis de fraude, crédito y recuperación y gestión de riesgo.
- Un modelo de LTV tiene tres componentes:
  - Valor del cliente en el tiempo (que tan frecuente se ordena)
  - la longitud de servicio del cliente (cuánto tiempo se tiene al cliente)
  - el valor de las órdenes (cuánto ingreso se obtiene del cliente)

  Cada componente se puede calcular por separado o en conjunto.
- Cuando se modela el LTV en el contexto de una campaña de retención, se tiene que calcular el LTV antes y después para medir el esfuerzo que se realiza.
- Los conceptos asociados al modelo son de la siguiente manera:

Ejemplo: Retención de clientes II

- T es el tiempo en que un cliente abandona la subscripción
- $S(t)$ es la probabilidad de que un cliente no haya abandonado la subscripción al tiempo t.
- probabilidad de que un cliente abandonará en el tiempo t: $\lambda(t)$.

- Los siguientes datos de Kaggle, que a su vez toma de IBM son sobre una compañía llamada Telco que ofrece servicios de comunicación. Son datos en donde cada renglón representa un cliente, y cada columna tiene atributos del cliente. Los datos incluyen información sobre:
  - Clientes que dejaron el servicio el último mes (*Churn*)
  - Servicios que el cliente tenía subscritos (teléfono, lineas múltiples, internet, seguridad, backup, protección de equipos, soporte técnico y streaming películas y TV)
  - Información de la cuenta del cliente (cuánto ha sido cliente, contrato, método de pago, estados de cuenta en línea, cargos mensuales y cargos totales)
  - Información demográfica de los clientes (sexo, rango de edad y si tienen pareja y dependientes)

Ejemplo: Retención de clientes III

```r
datos <- read.csv("../data/Churn/WA_Fn-UseC_-Telco-Customer-Churn.csv",header = T)
str(datos)
```

```
'data.frame':	7043 obs. of  21 variables:
 $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr  "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr  "Yes" "No" "No" "No" ...
 $ Dependents      : chr  "No" "No" "No" "No" ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr  "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr  "No" "No" "No" "No" ...
 $ StreamingMovies : chr  "No" "No" "No" "No" ...
 $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr  "No" "No" "Yes" "No" ...
```

```r
datos$Churn <- ifelse(datos$Churn == "Yes",1,0)  #Convertimos a dummy
```

- El número de clientes que abandonan en los datos disponibles

```
table(datos$Churn)


   0    1
5174 1869
```

Y la duración promedio de los clientes es

```
datos %>%
      group_by(Churn) %>%
      summarise(promedio = mean(tenure)) # el promedio es en meses

# A tibble: 2 x 2
  Churn promedio
  <dbl>    <dbl>
1     0     37.6
2     1     18.0
```

- Podemos estimar la curva de sobrevivencia usando el modelo no paramétrico de Kaplan-Meier. La variable 'tenure' indica el número total de meses que el cliente ha estado con la compañía.

```
KM <- survfit(Surv(tenure, Churn) ~ 1, data = datos)
KM

Call: survfit(formula = Surv(tenure, Churn) ~ 1, data = datos)

        n events median 0.95LCL 0.95UCL
[1,] 7043   1869     NA      NA      NA
```

Ejemplo: Retención de clientes V

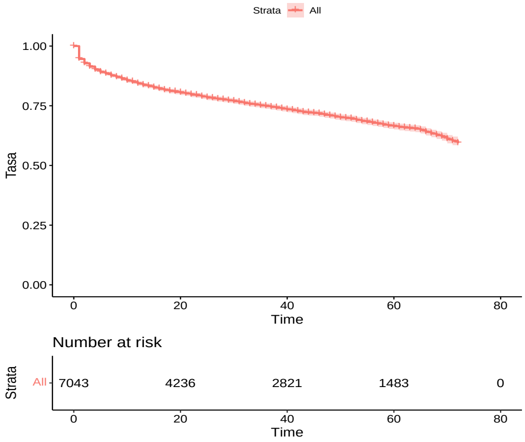
- Se devuelve el número de observaciones por curva, el número de eventos, la supervivencia mediana (aquí es NA, hay un caso con NA) y los intervalos de confianza para la mediana. Los tiempos de supervivencia mediana representan el tiempo en el que la probabilidad de supervivencia $S(t) = 0.5$.
- Podemos obtener más detalle con `summary`

```
sKM <- summary(KM)
str(sKM)
```

```
List of 18
 $ n            : int 7043
 $ time         : num [1:72] 1 2 3 4 5 6 7 8 9 10 ...
 $ n.risk       : num [1:72] 7032 6419 6181 5981 5805 ...
 $ n.event      : num [1:72] 380 123 94 83 64 40 51 42 46 45 ...
 $ n.censor     : num [1:72] 244 115 106 93 69 70 80 81 73 71 ...
 $ surv         : num [1:72] 0.946 0.928 0.914 0.901 0.891 ...
 $ std.err      : num [1:72] 0.0027 0.0031 0.00338 0.00361 0.00377 ...
 $ cumhaz       : num [1:72] 0.054 0.0732 0.0884 0.1023 0.1133 ...
 $ std.chaz     : num [1:72] 0.00277 0.00327 0.00362 0.00393 0.00417 ...
 $ type         : chr "right"
 $ logse        : logi TRUE
 $ conf.int     : num 0.95
 $ conf.type    : chr "log"
 $ lower        : num [1:72] 0.941 0.922 0.907 0.894 0.884 ...
 $ upper        : num [1:72] 0.951 0.934 0.92 0.908 0.899 ...
 $ call         : language survfit(formula = Surv(tenure, Churn) ~ 1, data = datos)
 $ table        : Named num [1:9] 7043 7043 7043 1869 54.5 ...
  ..- attr(*, "names")= chr [1:9] "records" "n.max" "n.start" "events" ...
 $ rmean.endtime: num 72
 - attr(*, "class")= chr "summary.survfit"
```

Ejemplo: Retención de clientes VI

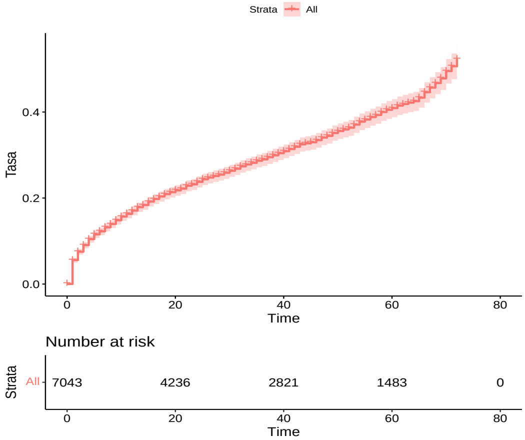
```r
ggsurvplot(KM, risk.table = T,
           main = "Tiempo hasta baja del cliente", ylab = "Tasa")
```



Ejemplo: Retención de clientes VII

- La gráfica de la función de riesgo acumulada se obtiene como

```
ggsurvplot(KM, risk.table = T,
           main = "Tiempo hasta baja del cliente", ylab = "Tasa", fun="cumhaz")
```

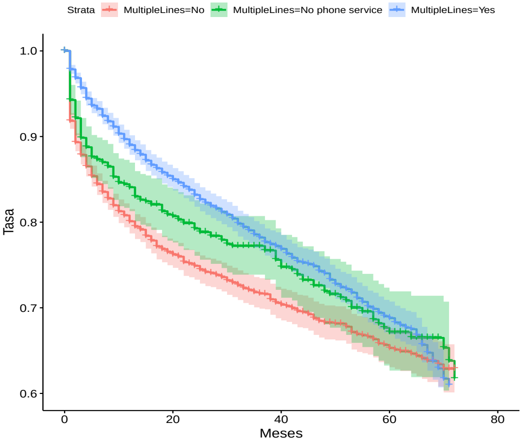


- Podemos considerar grupos y ver cómo se ve la permanencia en cada grupo. Por ejemplo, la variable ‘MultipleLines‘ indica 3 grupos: los que sí, los que no y los que no tienen el servicio telefonico

```
KM2 <- survfit(Surv(tenure, Churn) ~ MultipleLines, data = datos)
ggsurvplot(KM2, conf.int = T, main = "Tiempo hasta baja del contrato", xlab = "Meses", ylim = c(0.6, 1),
           ylab = "Tasa")
```

```
Warning: Removed 1 row(s) containing missing values (geom_path).
Warning: Removed 1 rows containing missing values (geom_point).
Warning: Removed 1 row(s) containing missing values (geom_path).
Warning: Removed 1 rows containing missing values (geom_point).
```

Ejemplo: Retención de clientes IX



Aquí podemos ver que los clientes que tienen múltiples lineas tiende a ser más fieles al menos los primeros 24 meses. Sin embargo, a partir de los 40 meses, ya no hay tanta distinción entre los grupos.

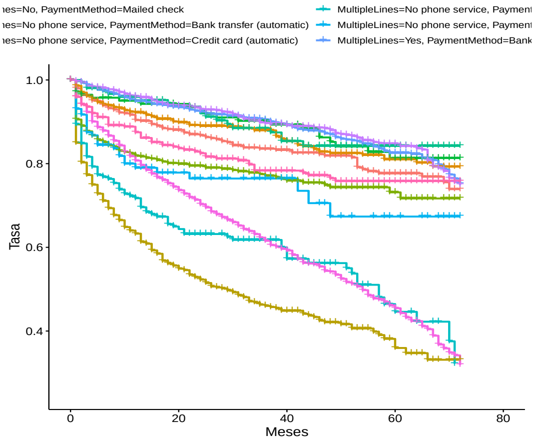
Ejemplo: Retención de clientes X

- Podemos añadir más variables e ir identificando quiénes son los grupos que son los que abandonan más rápido

```
KM3 <- survfit(Surv(tenure, Churn) ~ MultipleLines + PaymentMethod, data = datos)
ggsurvplot(KM3, conf.int = F, main = "Tiempo hasta baja del contrato", xlab = "Meses",
           ylab = "Tasa", ylim = c(0.25,1))
```

```
Warning: Removed 1 row(s) containing missing values (geom_path).
Warning: Removed 1 rows containing missing values (geom_point).
Warning: Removed 1 row(s) containing missing values (geom_path).
Warning: Removed 1 rows containing missing values (geom_point).
```

Ejemplo: Retención de clientes XI



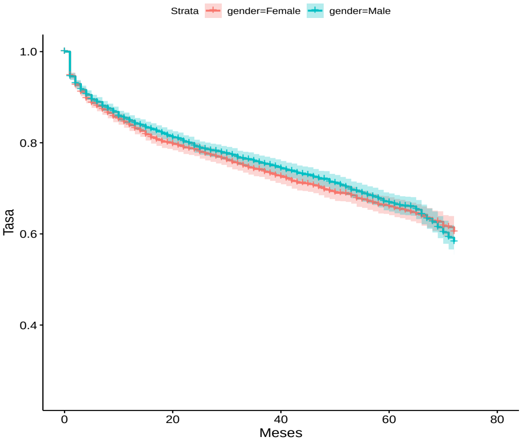
- Para hombres y mujeres:

```
KM4 <- survfit(Surv(tenure, Churn) ~ gender, data = datos)
print(KM4)  # Los NA indican que sólo unos cuantos clientes abandonaron durante el tiempo de observación
```

```
Call: survfit(formula = Surv(tenure, Churn) ~ gender, data = datos)

                 n events median 0.95LCL 0.95UCL
gender=Female 3488    939     NA      NA      NA
gender=Male   3555    930     NA      NA      NA
```

```
ggsurvplot(KM4, conf.int = T, main = "Tiempo hasta baja del contrato", xlab = "Meses",
           ylab = "Tasa", ylim = c(0.25,1))  # N se ven diferencias significativas en los grupos
```

Ejemplo: Retención de clientes XIII



El modelo de riesgos proporcionales de Cox se utiliza para cuantificar el riesgo de observar el evento de interés durante el periodo de observación. También se utilizar para evaluar simultáneamente el efecto de varias covariadas en el tiempo de supervivencia.
En este contexto, la función de riesgo se escribe como

$$\lambda(t) = \lambda(t|X) = \lambda_0(t)e^{\beta' X} = \lambda_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}$$

donde X_i es un vector de predictores y β es un vector de coeficientes de regresión. El riesgo base $\lambda_0(t)$ es el riesgo cuando todas los predictores son 0. Este valor en realidad no se estima, como veremos a continuación.
El modelo supone que no hay eventos empatados.
Razón de riesgos
El modelo de riesgos proporcionales estima la razón de los valores de los riesgos entre dos niveles, digamos X y X^*, con diferentes valores en los predictores. Se estima como:

$$HR = \frac{\lambda_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}}{\lambda_0(t)e^{\sum_{i=1}^{p} \beta_i X_i^*}} = e^{\sum_{i=1}^{p} \beta_i (X_i - X_i^*)}$$

Noten que la razón de riesgos no depende de t. Esto quiere decir que las curvas de sobrevivencia para dos grupos deben tener funciones de riesgo que son proporcionales para todos los valores de t y adicionalmente el cociente de riesgos no varía con el tiempo. Gráficamente, si las funciones de riesgo se cruzan, el supuesto es violado.

Modelo de Riesgos proporcionales de Cox III

```
modelo <- coxph(Surv(tenure, Churn) ~ gender + Dependents + MultipleLines + Partner + Contract, data = datos)
summary(modelo)

Call:
coxph(formula = Surv(tenure, Churn) ~ gender + Dependents + MultipleLines +
    Partner + Contract, data = datos)

  n= 7043, number of events= 1869

                                 coef exp(coef) se(coef)       z Pr(>|z|)
genderMale                   -0.04176   0.95910  0.04631  -0.902 0.367157
DependentsYes                -0.21228   0.80874  0.06659  -3.188 0.001432 **
MultipleLinesNo phone service -0.30853   0.73452  0.08463  -3.646 0.000266 ***
MultipleLinesYes             -0.34988   0.70477  0.05083  -6.883 5.87e-12 ***
PartnerYes                   -0.52912   0.58912  0.05408  -9.785  < 2e-16 ***
ContractOne year             -2.15960   0.11537  0.08408 -25.684  < 2e-16 ***
ContractTwo year             -4.14327   0.01587  0.15788 -26.243  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                              exp(coef) exp(-coef) lower .95 upper .95
genderMale                      0.95910      1.043   0.87587   1.05023
DependentsYes                   0.80874      1.236   0.70979   0.92148
MultipleLinesNo phone service   0.73452      1.361   0.62226   0.86704
MultipleLinesYes                0.70477      1.419   0.63794   0.77861
PartnerYes                      0.58912      1.697   0.52988   0.65499
ContractOne year                0.11537      8.668   0.09784   0.13604
ContractTwo year                0.01587     63.008   0.01165   0.02163

Concordance= 0.824  (se = 0.004 )
Likelihood ratio test= 2839  on 7 df,   p=<2e-16
Wald test            = 1466  on 7 df,   p=<2e-16
Score (logrank) test = 2512  on 7 df,   p=<2e-16
```