Universidad Nacional Autónoma de México

Facultad de Ciencias

Análisis de supervivencia con eventos recurrentes

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuaria

PRESENTA:

Reyes Solleiro Ana Sofía

TUTORA

Dra. Lizbeth Naranjo Albarrán

Ciudad Universitaria, CD. MX., noviembre 2018.

1.Datos del alumno.
Reyes
Solleiro
Ana Sofía
044-55-31-99-68-79
Universidad Nacional Autónoma de
México
Facultad de Ciencias
Actuaría
310263140

2.Datos del Tutor
Dra.
Lizbeth
Naranjo
Albarrán

3.Datos del sinodal 1
Dr.
Ricardo
Ramírez
Aldana

4.Datos del sinodal 2
M en C.
Salvador
Zamora
Muñoz

5.Datos del sinodal 3
Act.
Jaime
Vázquez
Alamilla

6.Datos del sinodal 4
Act.
Ángel Manuel
Godoy
Aguilar

7.Datos del trabajo escrito.
Análisis de supervivencia con eventos recurrentes
86p
2018

DEDICATORIA

A mi abuela por haber sido un gran ejemplo de fortaleza, tenacidad y ganas de vivir.

Agradecimientos

A Ángel y Elizabeth por darme las mejores herramientas para construir este camino, por darme las bases para siempre ser mejor persona. Sin ustedes nada de esto sería posible ¡LOS AMO!. Gracias por siempre impulsarme y motivarme a perseguir mis sueños y nunca rendirme.

A Pablo, por ser un gran apoyo, un gran hermano y por todo el amor incondicional.

A mi alma mater la UNAM, a la facultad de Ciencias por haberme dado los mejores años de mi vida, la mejor educación y grandes amigos.

A la Dra.Lizbeth Naranjo y a mis sinodales por sus valiosas correciones y apoyo durante el desarrollo de esta tesis.

A Contreras por ser un ejemplo de dedicación y entrega, por el apoyo incondicional y paciencia.

Índice general

Introducción

El análisis de supervivencia es una rama de la estadística que surge de la necesidad de conocer la distribución de los tiempos de vida y muerte de los individuos en algún estudio. Es además una de las ramas que tiene mayores aplicaciones en múltiples campos de estudio tales como la demografía, ciencia actuarial, medicina, ingeniería, entre muchos otros.

Hasta hace unos años únicamente se estudiaba el tiempo de supervivencia al primer evento, esto es, a la primera vez que a un individuo en un estudio le ocurriera el evento de interés, por ejemplo, en medicina, la primera vez que un paciente presentó bronquitis. En ingeniería el momento en el que una máquina falló. Pero actualmente no es importante únicamente la primera vez, sino cuántas veces o en qué lapsos sucedieron los eventos bajo estudio.

Es por esto que surge el análisis de supervivencia para eventos recurrentes, con el cual se busca modelar los distintos tiempos de falla para así obtener tasas, predicciones sobre los distintos tiempos de falla que puede presentar un individuo. A pesar de ser un tema interesante, al día de hoy encontrar información acerca del tema no es una tarea sencilla.

El objetivo principal de esta tesis es dar a conocer los distintos modelos que se han desarrollado para poder trabajar con datos del tipo recurrente. La tesis está divida en 2 capítulos; el primero retoma los conceptos básicos del análisis de supervivencia necesarios para entender los modelos de eventos recurrentes, en el segundo capítulo se desarrolla el tema de análisis de supervivencia para eventos recurrentes, poniendo principal atención a los modelos semiparamétricos: Andersen-Gill, Pretince-Williams-Peterson, y el modelo marginal de Wei-Lin-Weissfeld. A través del desarrollo de la tesis se van ejemplificando los temas con bases de datos precargadas en el software estadístico R. Finalmente, para el caso de eventos recurrentes se toma la base de datos de cáncer de vejiga.

Capítulo 1

Análisis de Supervivencia

1.1. Definición

El análisis de supervivencia surge con el fin de monitorear el tiempo de ocurrencia de algún evento de interés durante un tiempo definido de observación. Los estudios de supervivencia son parte de los estudios longitudinales. El mayor número de aplicaciones puede observarse en las ciencias de la salud, sin embargo no es el único campo de estudio. Cuando se está llevando a cabo un estudio en medicina se espera determinar el tiempo de muerte de un paciente. También resulta de utilidad medir el tiempo de falla de una máquina, la permanencia de un trabajador en cierto puesto, entre otros.

En el análisis de supervivencia, la variable de interés es el tiempo de falla o muerte. Como no se sabe en qué momento ocurrirá dicha falla, este tiempo es una variable aleatoria. El tiempo de supervivencia está definido como el tiempo que transcurre desde el inicio de un estudio o diagnóstico (en caso de alguna enfermedad) hasta el tiempo en el que ocurre el evento de interés (falla).

Es importante señalar que para poder medir el tiempo de supervivencia, el tiempo de inicio y falla del estudio deben estar bien definidos. El tiempo de inicio no es necesariamente el mismo para todos los individuos del estudio.

1.1.1. Censura

Cuando se está realizando un estudio, existen factores que complican que el estudio se lleve a cabo exitosamente. Uno de los principales problemas es la perdida de información. En ocasiones los individuos salen del estudio sin razón aparente lo que conlleva a tener información parcial sobre su tiempo de falla, a esta información parcial se le llama censura.

En especial se tienen 3 tipos de censura.

Censura Tipo I Usualmente los estudios tienen un tiempo establecido para su realización, durante ese tiempo deben llevarse a cabo las observaciones de los individuos, por lo tanto, los individuos que no hayan presentado falla dentro del tiempo de estudio se denominarán observaciones censuradas.

Censura Tipo II En este tipo de censura existe una dependencia entre el tamaño de muestra (n) y el número de fallas requeridas por el investigador, únicamente serán tomadas en cuenta las primeras m fallas, donde $m \leq n$. En este caso, el investigador decide prolongar la observación de los individuos en el estudio hasta que ocurran m fallas de n posibles.

Censura Tipo III "censura aleatoria". Ésta sucede cuando los sujetos salen del estudio sin haber presentado alguna falla y sin control del experimentador.

Existe además otra clasificación sobre la censura dependiendo el momento en que ocurre:

Censura por la derecha La censura por la derecha ocurre cuando el Seguimiento u observación termina y el sujeto no presentó la falla, o cuando el sujeto desaparece del estudio. De haber ocurrido la falla (no observada), ésta se presentaría después del tiempo de censura observado. En la figura 1.1 se tienen 4 sujetos en el estudio, únicamente el sujeto C presenta un evento al tiempo $t = 2$ mientras que los otros 3 están censurados por la derecha.

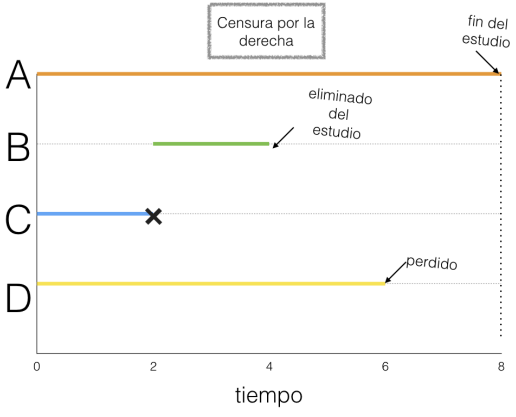

Figura 1.1: Censura por la derecha.

Censura por la izquierda La censura por la izquierda se da cuando la falla ocurre

antes de un tiempo específico observado, es decir, el sujeto presenta la falla antes de ingresar al estudio. Por ejemplo si se está realizando el Seguimiento de una persona para saber en qué momento resulta positiva en VIH, se tomaría en cuenta el momento a partir de que se realice una prueba. En el caso de la figura 1.2 los sujetos B y C presentaron el virus antes de que la prueba fuera realizada, donde la censura está representada como $\circ$ y la falla como $\times$.

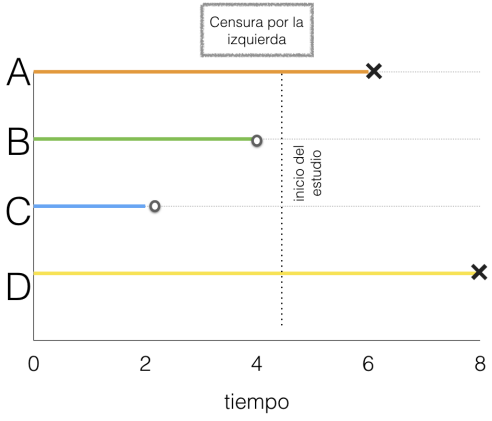

Figura 1.2: Censura por la izquierda.

Censura por intervalo Se dice que un sujeto tiene censura por intervalo cuando el evento de interés ocurre en algún momento del estudio entre dos tiempos. Por ejemplo, en la figura 1.3 se muestra el Seguimiento de un estudio para detectar VIH, para esto se harán 2 pruebas al tiempo t_1 y al t_2, sin embargo el sujeto en cuestión contrae la enfermedad entre los dos tiempos, sin que se sepa el momento exacto en el que ocurre la falla.

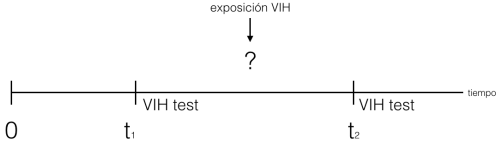

Figura 1.3: Censura por intervalo.

1.2. Funciones Involucradas en el Análisis de Supervivencia

El tiempo de supervivencia T es una variable aleatoria, por lo que se puede obtener su función de distribución. La variable T está definida siempre al inicio como $t = 0$ y el interés se encuentra en saber si el sujeto sobrevivirá más allá del tiempo t.

La distribución de los tiempos de supervivencia está caracterizada por 3 funciones que se definen en las siguientes subsecciones, estas funciones están relacionadas entre si, por lo que al obtener una es posible deducir las otras.

1.2.1. Función de Supervivencia

A la probabilidad de sobrevivir se le denota como $S(t)$ y se define como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t) \end{aligned}$$

La función de supervivencia al tiempo $t = 0$ es uno, es decir, $S(0) = 1$. $S(t)$ es una función no creciente pues es claro que si $t_1 < t_2$, entonces $S(t_1) \geq S(t_2)$, además el $\lim_{t\to\infty} S(t) = 0$.

1.2.2. Relación entre Función de Supervivencia y Función de Densidad

La función de densidad de probabilidad puede expresarse en términos de la función de supervivencia. La densidad se conoce además como tasa de falla incondicional.

$$\begin{aligned} S(t) &= P(T \geq t), \quad t \geq 0 \\ &= \int_t^{\infty} f_T(u)du \\ &= 1 - \int_0^t f_T(u)du \end{aligned} \tag{1.1}$$

donde f_T es la función de densidad de la variable aleatoria T, por el teorema fundamental del cálculo la ecuación (1.1) puede escribirse como

$$f_T(t) = -\frac{dS(t)}{dt} \tag{1.2}$$

1.2.3. Función de Riesgo

La función de riesgo $h(t)$ expresa la tasa de falla condicional. Esta tasa es la probabilidad de que un individuo que ha sobrevivido hasta t, falle en un intervalo muy pequeño $(t, t + \Delta t)$

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{1.3}$$

Calculando el límite,

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0^+} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t P(T \geq t)} && (1.4) \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t} && (1.5) \\ h(t) &= \frac{f_T(t)}{S(t)} = \frac{f_T(t)}{1 - F(t)} \\ &= -\frac{d}{dt} \log S(t) && (1.6) \end{aligned}$$

Si despejamos de la ecuación (1.6) a $S(t)$ obtenemos:

$$\begin{aligned} S(t) &= \exp\left\{ -\int_0^t h(u) du \right\} \\ &= \exp\{-H(t)\} \end{aligned} \tag{1.7}$$

donde $H(t) = \int_0^t h(u) du$ se conoce como la función de riesgo acumulado.

En la figura 1.4 se muestra un ejemplo de la función de riesgo, densidad y supervivencia de una variable aleatoria exponencial.

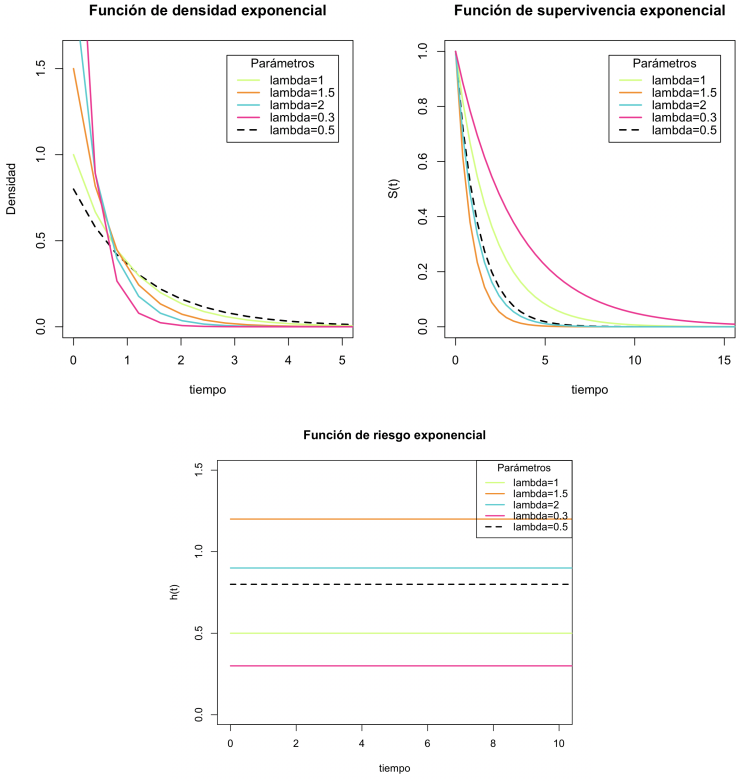

Figura 1.4: Funciones de una variable aleatoria T con distribución exponencial.

1.3. Estimador Kaplan-Meier

El estimador de Kaplan-Meier (KM) también conocido como estimador producto límite es el estimador máximo verosímil no paramétrico que busca estimar la función de supervivencia de una población. La probabilidad de supervivencia al tiempo $t_{(f)}$ donde f representa el f-ésimo tiempo ordenado, para el estimador KM, está dada por la ecuación (1.8).

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \times \hat{P}(T > t_{(f)} | T \geq t_{(f)}) \tag{1.8}$$

$$\hat{S}(t_{(f-1)}) = \prod_{i=1}^{f-1} \hat{P}(T > t_{(i)} | T \geq t_{(i)}) \tag{1.9}$$

En términos de los datos el estimador de la probabilidad de la ecuación (1.8) $\hat{P}(T > t_{(f)}|T \geq t_{(f)})$ se obtiene como $\frac{n_f - d_f}{n_f}$.

Esta ecuación indica la probabilidad de sobrevivir dado que el individuo sobrevivió al tiempo $t_{(f-1)}$ por la probabilidad condicional de vivir más allá del tiempo $t_{(f)}$ dado que vivió al menos hasta $t_{(f)}$.

Se sustituye este valor en la ecuación (1.8) y se obtiene la supervivencia al tiempo t_f,

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \times \frac{n_f - d_f}{n_f} \tag{1.10}$$

donde n_f indica el número de individuos en riesgo al inicio del estudio, d_f indica las pérdidas en el intervalo.

Demostración. $P(A \cap B) = P(A) \times P(B|A)$.

Sea $A = T \geq t_{(f)}$ y $B = T > t_{(f)} \implies P(A \cap B) = P(B)$ pues $A \subset B$ y $P(T > t_{(f)}) = S(t_{(f)})$

$P(A) = P(T \geq t_{(f)})$ puede ser visto como la probabilidad de que el sujeto haya sobrevivido al tiempo $t_{(f-1)}$ por lo que

$$\begin{aligned} P(A) &= P(T > t_{(f-1)}) = S(t_{(f-1)}) \\ P(B|A) &= P(T > t_{(f)}|T \geq t_{(f)}) \end{aligned}$$

de aquí se obtiene (1.8) . □

Para calcular los intervalos de confianza, se necesita calcular la varianza del estimador Kaplan-Meier. La varianza se puede calcular a través de la fórmula de Greenwood,

$$\widehat{Var}[\widehat{S}(t)] = \left(\widehat{S}(t)\right)^2 \sum_{f=1}^{k} \frac{d_f}{n_f(n_f - d_f)} \tag{1.11}$$

Donde k corresponde al k-ésimo tiempo ordenado. La demostración para obtener la varianza (1.11) puede consultarse en [Collet, 2015]. Una vez obtenida la varianza, es posible calcular los intervalos de confianza. Estos intervalos no siempre están entre 0 y 1 por lo que se recomienda hacer una transformación logaritmo o logit de los datos. La fórmula para calcular los intervalos al 95 % de confianza se muestra en (1.12) donde $Z_{1-\alpha/2}$ representa el percentil $1 - \alpha/2$ de una Normal.

$$\widehat{S}(t) \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}[\widehat{S}(t)]} \tag{1.12}$$

Para ejemplificar el estimador KM se utilizarán los datos de la base *leukemia* de la paquetería survival en R [R Core Team, 2017]. Al tiempo $t = 0$, $S(0) = 1$. En la tabla

1.5 se muestran los datos que se utilizarán para el ejemplo, donde el símbolo + indica que el tiempo está censurado.

Datos											
9	13	13+	18	23	28+	31	34	45+	48	161+	5
5	8	8	12	16+	23	27	30	33	43	45	

Tabla 1.1: Datos Leukemia

Para que R reconozca los datos como datos de supervivencia se utiliza la función Surv de la librería survival. El primer parámetro que recibe es el tiempo de falla de los datos y con el segundo parámetro se le indica a la función cómo está identificada la ocurrencia de un evento, en este caso 1 significa falla y 0 censura. (Cabe recordar que el doble signo igual = representa igualdad, un solo signo igual = se utiliza para asignar un valor a la variable). La sentencia en R para calcular el estimador es la siguiente:

```
km1 <- survfit(Surv(time,status==1, data=leukemia)~1)
```

En la tabla 1.2 y en la figura 1.5 se muestra la estimación de la función de supervivencia de los datos de leukemia.

tiempo	n.riesgo	n.evento	n.censuras	supervivencia	std.err	inferior 95 % IC	superior 95 % IC
5	23	2	0	0.9130	0.0588	0.8049	1
8	21	2	0	0.8261	0.0790	0.6848	0.996
9	19	1	0	0.7826	0.0860	0.6310	0.971
12	18	1	0	0.7391	0.0916	0.5798	0.942
13	17	1	2	0.6957	0.0959	0.5309	0.912
18	14	1	0	0.6460	0.1011	0.4753	0.878
23	13	2	0	0.5466	0.1073	0.3721	0.803
27	11	1	1	0.4969	0.1084	0.3240	0.762
30	9	1	0	0.4417	0.1095	0.2717	0.718
31	8	1	0	0.3865	0.1089	0.2225	0.671
33	7	1	0	0.3313	0.1064	0.1765	0.622
34	6	1	0	0.2761	0.1020	0.1338	0.569
43	5	1	0	0.2208	0.0954	0.0947	0.515
45	4	1	1	0.1656	0.0860	0.0598	0.458
48	2	1	1	0.0828	0.0727	0.0148	0.462

Tabla 1.2: Estimador Kaplan-Meier

$$
\begin{aligned}
S(t_1) &= S(t_0) \times \frac{23-2}{23} \qquad \text{donde } t_1 = 5 \\
S(5) &= 1 \times \frac{21}{23} = 0.9130
\end{aligned}
$$

El intervalo de confianza superior al tiempo t_1 es mayor que 1, por lo que como se mencionó anteriormente habría que hacer una transformación.

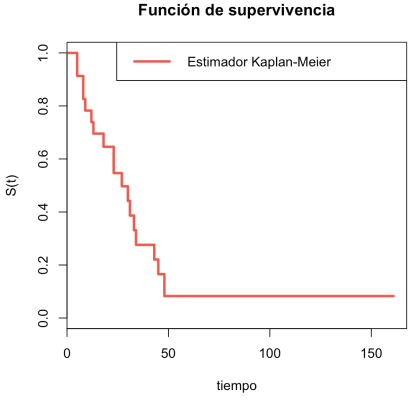

Figura 1.5: Gráfico estimador Kaplan-Meier.

1.4. Prueba Log-Rank

La prueba de Log- Rank es una prueba de hipótesis utilizada para probar si dos curvas de supervivencia son iguales. La prueba de log-rank se distribuye como una Ji-cuadrada para muestras grandes y la base de esta prueba es comparar los datos observados contra los esperados bajo la hipótesis nula (superviviencias iguales).

$$H_0 : S_1(t) = S_2(t) \ \forall \ t > 0 \text{ vs. } H_1 : S_1(t) \neq S_2(t) \text{ p.a } t > 0$$

En la base de datos de leucemia se tienen 2 grupos, los que recibieron seguimiento en su tratamiento y los que no. Para ejemplificar la prueba de Log-Rank se probará si las curvas de supervivencia iguales para ambos grupos. En la tabla 1.5 se muestran los datos para realizar el ejemplo. Las curvas estimadas para cada grupo se encuentran en la figura 1.6.

Los eventos esperados se calculan de la siguiente forma, donde n_{if} con $i = 1, 2$ representa el conjunto en riesgo al tiempo f y d_{if} las fallas al tiempo f para el grupo i-ésimo.

$$\begin{aligned} e_{1f} &= \left(\frac{n_{1f}}{n_{1f} + n_{2f}}\right) \times (d_{1f} + d_{2f}) \\ e_{2f} &= \left(\frac{n_{2f}}{n_{1f} + n_{2f}}\right) \times (d_{1f} + d_{2f}) \end{aligned}$$

Para comparar 2 grupos, el estadístico de Log-Rank utiliza la suma de la diferencia entre los eventos esperados y observados sobre todos los tiempos de falla de uno de los 2

Tiempo	Estatus	Grupo	Tiempo	Estatus	Grupo
9	1	Seguimiento	5	1	Sin Seguimiento
13	1	Seguimiento	5	1	Sin Seguimiento
13	0	Seguimiento	8	1	Sin Seguimiento
18	1	Seguimiento	8	1	Sin Seguimiento
23	1	Seguimiento	12	1	Sin Seguimiento
28	0	Seguimiento	16	0	Sin Seguimiento
31	1	Seguimiento	23	1	Sin Seguimiento
34	1	Seguimiento	27	1	Sin Seguimiento
45	0	Seguimiento	30	1	Sin Seguimiento
48	1	Seguimiento	33	1	Sin Seguimiento
161	0	Seguimiento	43	1	Sin Seguimiento
			45	1	Sin Seguimiento

Tabla 1.3: Datos Leucemia para la prueba Log-Rank

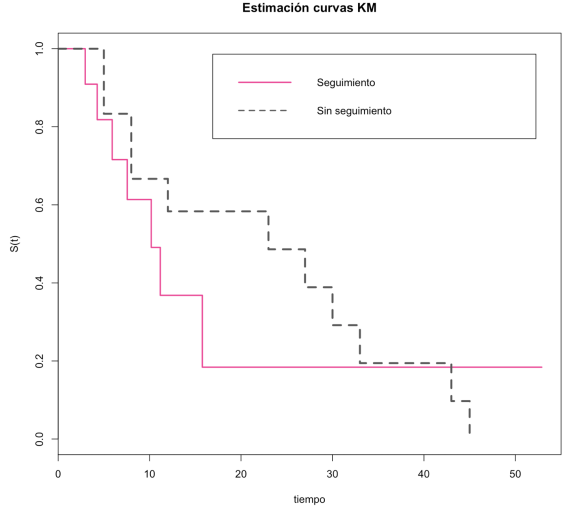

Figura 1.6: Curva de supervivencia estimada por grupo.

grupos al cuadrado entre la varianza de dicha diferencia. La varianza para ambos grupos, mostrada en la ecuación (1.13)se estima de suponer que la variable d_{1f} se distribuye

f	$t_{(f)}$	d_{1f}	d_{2f}	n_{1f}	n_{2f}
1	5	0	2	11	12
2	8	0	2	11	10
3	9	1	0	11	10
4	12	0	1	10	8
5	13	1	0	10	8
6	18	1	0	8	8
7	23	1	1	7	6
8	27	0	1	7	5
9	30	0	1	6	4
10	31	1	0	5	4
11	33	0	1	5	3
12	34	1	0	4	3
13	43	0	1	4	2
14	45	0	1	4	1
15	48	1	0	2	0
total		7	11		

Tabla 1.4: Datos ejemplo Log-rank

hipergeométrica.

$$
\begin{aligned}
Var(O_i - E_i) &= \sum_f \frac{n_{1f} n_{2f} (d_{1f} + d_{2f})(n_{1f} + n_{2f} - d_{1f} - d_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)} \qquad (1.13) \\
O_i - E_i &= \sum_f d_{if} - e_{if} \qquad i = 1, 2
\end{aligned}
$$

Estadístico Log-Rank
H_0 : no hay diferencia entre las curvas de supervivencia. vs H_1 : las curvas de supervivencia no son iguales

$$\frac{(O_i - E_i)^2}{Var(O_i - E_i)} \underset{\sim}{a} \chi^2_1 \qquad i = 1, 2 \tag{1.14}$$

El comando para hacer la prueba de Log-Rank en R es el siguiente, es posible además realizar esta prueba para más de 2 grupos y para datos estratificados.

```
log_rank <- survdiff(Surv(time, status) ~ x,data=leukemia)
```

Tabla 1.5: Prueba Log-Rank para datos de Leukemia

	N	Observados	Esperados	$(O-E)^2/E$	$(O-E)^2/V$
x= Seguimiento	11	7	10.69	1.27	3.4
x= No Seguimiento	12	11	7.31	1.86	3.4
Chisq = 3.4 con 1 grado de libertad			p = 0.0653		

En la tabla 1.5 están los resultados de la prueba los cuales muestran que a un nivel de significancia de $\alpha = 0.05$, el p-value es mayor $= 0.0653 > \alpha$, por lo que no se tiene información suficiente para rechazar la hipótesis nula, esto es, que las curvas son iguales.

1.5. Estimador Nelson-Aalen

El principal objetivo de este estimador es estimar la función de riesgo, tomando en cuenta que es más sencillo estimar la función de riesgo acumulada.

$$H(t) = \int_0^t h(s)ds \tag{1.15}$$

Para poder estimar la función (1.15) es necesario utilizar los procesos $\overline{N(t)}$ y $\overline{Y(t)}$,

$$\overline{N(t)} = \sum_{i=1}^{n} N_i(t) \qquad \text{y} \qquad \overline{Y(t)} = \sum_{i=1}^{n} Y_i(t) \tag{1.16}$$

donde $N_i(t)$ cuenta el número de eventos observados en el tiempo $(0, t)$ para el individuo i-ésimo, y el proceso Y_i indica si el individuo i está en riesgo justo antes de t.

Se considera un intervalo de tiempo pequeño:

$$\begin{aligned} H(t + \Delta t) - H(t) &\approx H(t)\Delta t \\ H(t + \Delta t) - H(t) &= P(t \leq T \leq t + \Delta t | T \geq t) \end{aligned} \tag{1.17}$$

La ecuación (1.17) se puede estimar como $\frac{\bar{N}(t+\Delta t)-\bar{N}(t)}{\bar{Y}(s)}$, haciendo la suma de los intervalos pequeños que contienen a lo más 1 evento se obtiene el estimador de Nelson-Aalen:

$$\hat{H}(t) = \sum_{i:t_i \leq t} \frac{\Delta \bar{N}(t_i)}{\bar{Y}(t_i)} \tag{1.18}$$

Ejemplo

Se tienen 70 tiempos de falla de ventiladores de diesel. El objetivo del estudio es saber si hay que reemplazar los ventiladores que están funcionando por ventiladores nuevos de mayor calidad para evitar futuras fallas. El problema es identificar si la tasa de falla disminuye con el tiempo, ya que es posible que las primeras fallas eliminen los ventiladores más débiles y las fallas futuras sean tolerables. [1]

4.5	4.6+	11.5	11.5	15.6+	16.0	16.6+	18.5+	18.5+	18.5+	18.5+	18.5+	20.3+	20.3+	20.3+
20.7	20.7	20.8	22.0+	30.0+	30.0+	30.0+	30.0+	31.0	32.0+	34.5	37.5+	37.5+	41.5+	41.5+
41.5+	41.5+	43.0+	43.0+	43.0+	43.0+	46.0	48.5+	48.5+	48.5+	48.5+	50.0+	50.0+	50.0+	61.0
61.0+	61.0+	61.0+	63.0+	64.5+	64.5+	67.0+	74.5+	78.0+	78.0+	81.0+	81.0+	82.0+	85.0+	85.0+
85.0+	87.5+	87.5	87.5+	94.0+	99.0+	101.0+	101.0+	101.0+	115.0+					

Figura 1.7: Tiempos falla de miles de horas de funcionamiento de ventiladores

La figura 1.7 muestra los tiempos de falla ordenados por ocurrencia de los ventiladores, donde + significa censura. El comando en R es:

```
Surv(survt,status==1,data=ventiladores)
```

Se denota T_i^* al tiempo ocurrido hasta la falla del ventilador i-ésimo, $i = 1, ..., 70$. Se supone que las T_i^* son variables independientes e idénticamente distribuidas. Por otra parte C_i^* se refiere al tiempo de censura, por lo que la variable T_i se define como $T_i = min(T_i^*, C_i^*)$

En términos de procesos de conteo, $Y_i(t) = I\{T_i \geq t\}$, es la función indicadora de que el ventilador i-ésimo aún está en observación al tiempo t, la cual es 1 al tiempo $t = 0$, ya que es cuando el ventilador se pone a funcionar hasta que se censura u ocurre la falla. $N_i(t)$ es el número de fallas que ha presentado el ventilador i-ésimo, esta variable será 0 hasta que ocurra la primera falla. Para el primer ventilador, $Y_1(t) = 1$ en $t \leq 4.5$ y 0 después, $N_1(t) = 0$ en $t < 4.5$ y 1 después.

Para obtener el estimador Nelson-Aalen en R se debe calcular la suma acumulada de la siguiente manera: $cumsum(n.risk/n.event)$ con lo que se obtienen los datos presentados en la tabla 1.7.

Es claro que al tiempo $t = 3$ arbitrario , $H(3) = 0$ pues no hubo ninguna falla en $(0, 3]$ sino hasta $t = 4.5$. Al tiempo $t_1 = 4.5$ la curva aumenta en (# fallas al tiempo 4.5

[1] Fuente: [Therneau and Grambsch, 2000]

Tabla 1.6: Datos de tiempos de falla.

tiempo	n.riesgo	n.evento	supervivencia	std.err	inferior 95 % IC	superior 95 % CI
4.5	70	1	0.9857	0.0141	0.9583	1
11.5	68	2	0.9567	0.0244	0.9099	1
16	65	1	0.9420	0.0281	0.8884	0.9988
20.7	55	2	0.907	0.0360	0.8397	0.9812
20.8	53	1	0.8906	0.0392	0.8169	0.9709
31	47	1	0.8716	0.0427	0.7917	0.9596
34.5	45	1	0.8523	0.0459	0.7667	0.9473
46	34	1	0.8272	0.0509	0.7330	0.9334
61	26	1	0.7954	0.0581	0.6892	0.9178
87.5	9	1	0.7070	0.0980	0.5387	0.9278

Tabla 1.7: Estimador Nelson-Aalen.

Tiempo	Riesgo	$\hat{H}(t)$
4.5	1 /70	0.0142
11.5	2/68	0.0436
16.0	1/65	0.0590
20.7	2/55	0.0954
20.8	1/53	0.1143
31	1/47	0.1355
34.5	1/45	0.1578
46	1/34	0.1872
61	1/26	0.2256
87.5	1/9	0.3367

/ # ventiladores en riesgo al tiempo 4.5)= 1/70, al tiempo $t_3 = 16$ tiene un incremento de 1/65, y así sucesivamente hasta el $t_{10} = 87.5$, con lo que para cualquier $t \geq 87.5$, se tendrá el estimador:

$$\begin{aligned}\hat{H}(t) &= \frac{\#\text{ fallas en } 4.5}{\#\text{ ventiladores en riesgo a } 4.5} + ... + \frac{\#\text{ fallas en } 87.5}{\#\text{ ventiladores en riesgo a } 87.5} \\ &= \frac{1}{70} + \frac{2}{68} + ... + \frac{1}{9} \\ &= 0.3367\end{aligned}$$

El estimador $\hat{H}(t)$ estima el número promedio de fallas en $(0, t]$ por unidad en riesgo. Para el caso de este ejemplo, representa el número de fallas esperadas en t, tomando en cuenta que los ventiladores se pusieron a trabajar en $t = 0$.

El estimador de la función de riesgo acumulado obtenido por el método de Máxima Verosimilitud es $\hat{H}(t) = -\log \hat{S}(t)$, en la figura 1.8 se muestra la función de riesgo acumulado para los ventiladores obtenida mediante ambos estimadores. Se puede observar que ambos estimadores son muy similares. Por lo que para estimar la función podria utilizarse cualquiera de los dos, sin embargo el estimador Nelson-Aalen será de utilidad para el calculo de residuos.

A continuación se muestra el código en R para calcular y graficar el estimador Nelson

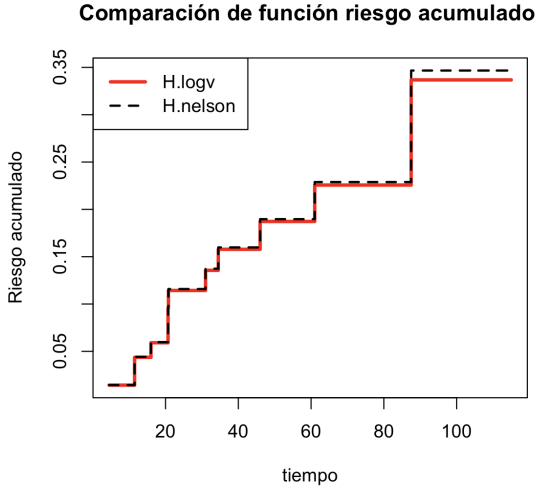

Figura 1.8: Estimador Nelson-Aalen para tiempos falla de ventiladores.

Aalen.

```
H.hat   <- -log(fit$surv) # EMV riesgo acumulado
H.hat   <- c(H.hat, tail(H.hat, 1))
#Estimador Nelson-Aalen
Lambda.hat   <- cumsum(fit$n.event / fit$n.risk)
Lambda.hat  <- c(Lambda.hat, tail(Lambda.hat, 1))
plot(c(fit$time, 100), Lambda.hat, xlab="time",
      ylab="cumulative hazard",
      main="Comparación función riesgo acumulado",
      ylim=range(c(H.hat, Lambda.hat)), type="s", col="blue")
points(c(fit$time, 100), H.hat, lty=2, type="s", col="red")
legend("topleft", legend=c("H.hat","Lambda.hat"),
       lty=1:2, col=c("red","blue"))
```

1.6. Modelo de Riesgos Proporcionales de Cox

El modelo de Cox fue desarrollado para poder determinar los efectos de las covariables en el riesgo instantáneo de la ocurrencia de un evento y en la supervivencia de los individuos bajo estudio.

El modelo de Cox se escribe usualmente en términos de la función de riesgo. Con esta función se obtiene el riesgo de un individuo al tiempo t, dado un vector de covariables X.

Las covariables se miden al inicio del estudio y se denotan como $X_1, X_2, X_3, \ldots, X_q$, el modelo que se presenta es el siguiente:

$$h(t|X) = h_0(t) \exp\left\{\sum_{j=1}^{q} \beta_j X_j\right\} \tag{1.19}$$

donde $h(t|X)$ es la función de riesgo instantáneo, dado el vector de covariables $X = (X_1, X_2, X_3, ..., X_q)$.

Como puede verse, la función de riesgo en la ecuación (1.19) está compuesta por dos partes: la primera es $h_0(t)$, ésta es la función de riesgo base, dependiente del tiempo pero independiente de las covariables. La segunda parte es la exponencial, que contiene el vector $\beta = (\beta_1, \beta_2, \beta_3, ..., \beta_q)$ de los coeficientes de las variables del vector de covariables. En caso de que las covariables dependan del tiempo, se necesita utilizar un modelo de Cox extendido, pues no se cumpliría la hipótesis de riesgos proporcionales. Si el vector β es nulo, las covariables no afectan el riesgo instantáneo de ocurrencia, lo que implica que $h(t|X) = h_0(t)$. La función de riesgo base $h_0(t)$ es una función desconocida.

El modelo de Cox es popular, ya que a pesar de que no se conoce la función base, es posible obtener estimadores de los coeficientes de regresión, tasas de riesgo y curvas de supervivencia ajustadas, esto implica que el modelo de Riesgos Proporcionales sea un modelo robusto, pues los resultados serán aproximados al modelo paramétrico.

Sería mucho más sencillo, utilizar modelos paramétricos, pero encontrar el modelo que mejor ajusta los datos es complicado, por esta razón se puede utilizar el modelo de Cox.

Otra razón por la que es cómodo utilizar un modelo de Cox es porque en la parte exponencial se obtendrán estimadores positivos, pues el soporte de la función exponencial es de 0 a infinito, y esto es conveniente ya que por definición la función de riesgo debe tomar valores no negativos.

El modelo de Cox asume que el riesgo instantáneo de dos individuos i y j distintos

es proporcional, esto es:

$$\frac{h(t|X_i)}{h(t|X_j)} = \frac{h_0(t)e^{\beta'(X_i)}}{h_0(t)e^{\beta'(X_j)}} = e^{\beta'(X_i - X_j)} \tag{1.20}$$

Como $e^{\beta'(X_i - X_j)} = \gamma$ es una constante que no depende del tiempo, se reescribe la ecuación (1.20) como

$$h(t|X_i) = \gamma h(t|X_j) \tag{1.21}$$

Retomando la ecuación (1.7), adaptada al modelo de Cox se tiene:

$$S(t|X) = e^{-\int_0^t h(s|X)ds} = e^{-H(t|X)} \tag{1.22}$$

donde $H(t|X)$ es la función de riesgo acumulado dado el vector de covariables X.
Por otra parte, si sustituimos el riesgo instantáneo de Cox en la función de riesgo acumulado se tiene:

$$\begin{aligned} H(t|X) &= \int_0^t h(s|X)ds \\ &= \int_0^t h_0(s)e^{\beta' X}ds = e^{\beta' X}H_0(t) \end{aligned} \tag{1.23}$$

así

$$S(t|X) = \left[e^{H_0(t)}\right]^{\theta} = S_0(t)^{\theta} \tag{1.24}$$

donde $\theta = e^{\beta' X}$.

Una vez que se desarrollaron las equivalencias, es de interés estimar los parámetros del modelo. Se busca estimar los coeficientes del vector β, en el caso del modelo de Cox, se utiliza el método de máxima verosimilitud, dado que está función no depende del riesgo h_0, se le conoce como verosimilitud parcial.

Supóngase que se tienen n individuos bajo estudio, r tiempos de falla distintos y $n-r$ tiempos censurados. Además se supondrá que en t solo muere un individuo y así no se tendrán tiempos repetidos.

Los r tiempos de falla ordenados son $t_{(1)} < t_{(2)} < t_{(3)} < ... < t_{(r)}$, $t_{(j)}$ es el j-ésimo tiempo de falla ordenado. El conjunto de individuos que está en riesgo al tiempo $t_{(j)}$ se denota como $R(t_{(j)})$ (conjunto en riesgo), y es el conjunto de individuos que están vivos y no censurados al tiempo exactamente anterior a $t_{(j)}$. Por otra parte, δ_i es una variable indicadora de censura, que será cero si el individuo i-ésimo está censurado, y uno en otro caso, para $i = 1, ..., n$.

La función de verosimilitud para el modelo de riesgos proporcionales de Cox está dada por la ecuación (1.25). En el siguiente ejemplo se mostrará como se obtiene esta ecuación.

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' x_i)}{\sum_{l \in R(t_{(i)})} \exp(\beta' x_l)} \right]^{\delta_i} \tag{1.25}$$

aplicando logaritmo a la ecuación (1.25) se obtiene, tomando en cuenta únicamente los datos no censurados

$$\log L(\beta) = \sum_{i=1}^{j} \left[\beta' x_i - \log \sum_{l \in R(t_i)} \exp(\beta' x_l) \right] \tag{1.26}$$

derivando con respecto a β.

$$\frac{\delta \log L(\beta)}{\delta \beta} = \sum_{i=1}^{j} \left[x_i - \frac{\sum_{l \in R(t_i)} \exp(\beta' x_l) x_l}{\sum_{l \in R(t_i)} \exp(\beta' x_l)} \right] \tag{1.27}$$

Se iguala la ecuación (1.27) a cero y se obtiene un sistema de ecuaciones no lineales con tantas incógnitas como covariables, para encontrar la solución se aplica algún método de aproximación numérica, como puede ser el método de Newton-Raphson.

Ejemplo

Se tiene 4 sujetos: Pedro, Pablo, Luis y Juan, ahora se consideran los datos de la figura 1.8: Se puede observar que Pedro al tiempo 3 está censurado, mientras que Pablo,

Tabla 1.8: Ejemplo: datos.

	tiempo	estatus	fumador
Pedro	3	0	1
Pablo	8	1	1
Juan	5	1	0
Luis	2	1	0

Juan y Luis no se censuran en ningún tiempo, Pablo y Pedro son fumadores. En la figura 1.9 se puede ver gráficamente la censura y los tiempos de falla de los individuos, donde $\circ$ representa la censura y $\times$ la falla.

Se va a considerar el modelo de Cox de riesgos proporcionales con un predictor que es fumador $h(t) = h_0(t) \exp^{\beta_1 \text{fumador}}$, entonces la función de riesgo para cada individuo se expresa tal como se muestra en la tabla 1.9 Se puede ver que al tiempo 2 todos los individuos están en riesgo por lo tanto, el conjunto en riesgo del primer término de la verosimilitud contiene a los 4 individuos, pero al tiempo 5, Pedro está censurado por lo que el conjunto en riesgo disminuirá a 2.

La verosimilitud del modelo estará dada por el producto de la verosimilitud de cada

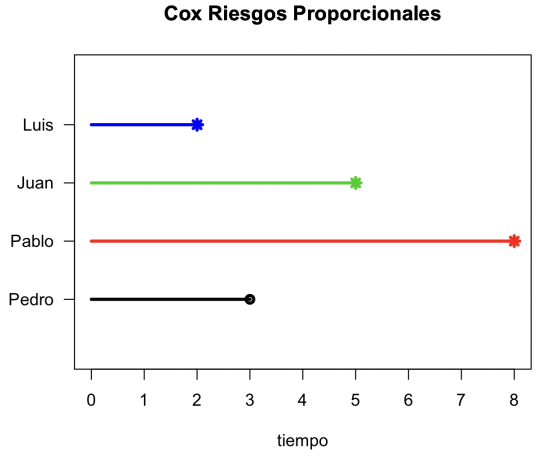

Figura 1.9: Tiempos de falla y censura de individuos en riesgo.

Tabla 1.9: Ejemplo: funciones de riesgo.

Sujeto	Función de riesgo
Pedro	$h_0(t)\exp^{\beta_1(1)}$
Pablo	$h_0(t)\exp^{\beta_1(1)}$
Juan	$h_0(t)\exp^{\beta_1(0)}$
Luis	$h_0(t)\exp^{\beta_1(0)}$

individuo no censurado.

$$
\begin{aligned}
L(\beta) &= \left[\frac{h_0(t)\exp^{\beta_1(0)}}{h_0(t)\exp^{\beta_1(0)}+h_0(t)\exp^{\beta_1(0)}+h_0(t)\exp^{\beta_1(1)}+h_0(t)\exp^{\beta_1(1)}}\right] \\
&\times \left[\frac{h_0(t)\exp^{\beta_1(0)}}{h_0(t)\exp^{\beta_1(0)}+h_0(t)\exp^{\beta_1(1)}}\right] \times \left[\frac{h_0(t)\exp^{\beta_1(1)}}{h_0(t)\exp^{\beta_1(1)}}\right] \\
L(\beta) &= \left[\frac{\exp^{\beta_1(0)}}{\exp^{\beta_1(0)}+\exp^{\beta_1(0)}+\exp^{\beta_1(1)}+\exp^{\beta_1(1)}}\right] \\
&\times \left[\frac{\exp^{\beta_1(0)}}{\exp^{\beta_1(0)}+\exp^{\beta_1(1)}}\right] \times \left[\frac{\exp^{\beta_1(1)}}{\exp^{\beta_1(1)}}\right] \\
L(\beta) &= \left[\frac{1}{2+2\exp^{\beta_1}}\right] \times \left[\frac{1}{1+\exp^{\beta_1}}\right] \times 1 \\
&= \left[\frac{1}{(2+2\exp^{\beta_1})(1+\exp^{\beta_1})}\right]
\end{aligned}
$$

Para encontrar el estimador de β se debe calcular el logaritmo de la función de verosimilitud $L(\beta)$.

$$\begin{aligned}\log L(\beta) &= \log\left[\frac{1}{(2+2\exp^{\beta_1})(1+\exp^{\beta_1})}\right] \\ &= \log(1) - \log(2+2\exp^{\beta_1}) - \log(1+\exp^{\beta_1})\end{aligned}$$

Entonces se estima β_1 de tal manera que $\log L(\beta)$ se maximice.

1.6.1. Pruebas de Hipótesis (Verosimilitud Parcial)

Existen 3 formas para probar la hipótesis nula $H_0 : \beta = 0$ dentro de la teoría de la verosimilitud: mediante la prueba de Wald, la prueba de puntaje (*score test*) y la prueba del cociente de verosimilitudes (*likelihood ratio test*). Estas tres pruebas, generalmente dan resultados similares.

Para llevar a cabo las pruebas es necesario contar con dos funciones: la primera es la derivada del logaritmo de la función de verosimilitud parcial (función score), cuya representación es la siguiente: $S(\beta) = L'(\beta)$, la segunda función es la menos derivada de la función score $I(\beta) = -S'(\beta) = L''(\beta)$. Si se sustituye a β por $\hat{\beta}$ se obtiene la matriz de información observada de Fisher. El cálculo de estos 3 estimadores a detalle se puede consultar en [Moore, 2016].

Prueba de Wald

La prueba de Wald es una de las pruebas más utilizadas. El estadístico de la prueba de Wald se obtiene a partir de dividir el estimador $\hat{\beta}$ entre el error estándar del estimador. El estadístico obtenido Z_w se asume que sigue una distribución normal asintótica.

$$Z_w = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

La hipótesis nula $H_0 : \beta = 0$ vs. $H_a : \beta \neq 0$ se rechaza si $|Z_w| > Z_{\alpha/2}$, donde $Z_{\alpha/2}$ representa el cuantil $\alpha/2$ de la distirbución normal.

Prueba de puntajes

La prueba de puntajes se obtiene de dividir la primera derivada de la función de log verosimilitud parcial, $S(\beta) = L'(\beta)$ y la varianza del estadístico score $I(\beta)$. Normalmente se evalúan ambas funciones en la hipótesis nula $\beta = 0$. Esta prueba es equivalente a la prueba de logrank. El estadístico de prueba Z_s se obtiene entonces como:

$$Z_s = \frac{S(\beta = 0)}{\sqrt{I(\beta = 0)}}$$

la hipótesis nula $H_0 : \beta = 0$ vs. $H_a : \beta \neq 0$ se rechaza si $|Z_s| > Z_{\alpha/2}$, donde $Z_{\alpha/2}$ representa el cuantil $\alpha/2$ de la distribución normal.

Prueba del cociente de verosimilitudes

La prueba del cociente de verosimilitudes utiliza la log verosimilitud calculada anteriormente, tomando el hecho de que $T = 2[L(\beta = \hat{\beta}) - L(\beta = 0)]$ se aproxima a una variable aleatoria con distribución Ji-cuadrado con 1 grado de libertad.

La hipótesis nula $H_0 : \beta = 0$ vs. $H_a : \beta \neq 0$ se rechaza si $T > \chi^2_{1,1-\alpha}$.
Si se quiere calcular estos tres estadísticos para el caso de múltiples parámetros la hipótesis nula a probar en todos los casos es H_0 : todas las β son iguales a cero vs. al menos una es distinta de 0. Todas las pruebas tienen una distribución Ji-cuadrada con p grados de libertad, donde p es el número de covariables en el modelo.

- Prueba del cociente de verosimilitudes:

$$2\{\log L(\hat{\beta}) - \log L(0)\}$$

- Prueba de Wald:

$$\hat{\beta}' I(\hat{\beta})\hat{\beta}$$

- Prueba de puntajes: en este caso

$$u'(0)I^{-1}(0)u(0)$$
$$u(0) = \frac{\partial \log L(0)}{\partial \beta_j}$$

En [R Core Team, 2017] la manera de obtener los 3 estadísticos es mediante la función de Cox, para realizar el ejemplo se utilizaron los datos de la base *leukemia* donde la variable x indica el grupo al que pertenecen los individuos. En este ejemplo se estudian los individuos que recibieron seguimiento en el tratamiento de leucemia, por lo que la variable x puede tomar solamente 2 valores: seguimiento (*maintained*) y sin seguimiento (*Nonmaintained*)

```
result.cox <- coxph(Surv(time,status) ~ x, data=leukemia)
summary(result.cox)
```

Tabla 1.10: Salida de R para pruebas de contraste de hipótesis

	coef	exp(coef)	se(coef)	z	$P(>\|z\|)$
xNonmaintained	0.9155	2.498	0.511	1.788	0.0737
Rsquare = 0.137					
Likelihood ratio test	= 3.38 on 1 df,	p = 0.06581			
Wald test	= 3.2 on 1 df,	p = 0.07371			
Score (logrank) test	= 3.42 on 1 df,	p = 0.06454			

Los resultados de las pruebas pueden consultarse en la tabla 1.10. A un nivel de significancia $\alpha = 0.05$ se puede ver que en las 3 pruebas se tiene un p-value mayor que α por lo que no se rechaza la hipótesis nula, es decir, $\beta = 0$.

1.7. Supuestos del Modelo de Cox

El modelo de riesgos proporcionales de Cox, supone que la razón de riesgos (HR) será constante a través del tiempo. La razón de riesgos está definida como la división del riesgo de un individuo i entre el riesgo de un individuo j distintos, que se diferencian por las covariables.

$$\widehat{HR} = \frac{\hat{h}(t, X_i)}{\hat{h}(t, X_j)} = \hat{\theta}$$

Para probar este supuesto se tienen 2 formas distintas de hacerlo: la prueba de bondad de ajuste y mediante las variables dependientes del tiempo, el método gráfico, por otra parte, sirve para tener una idea general de supuesto de riesgos proporcionales en el modelo. Se han desarrollado otras técnicas para la verificación de riesgos proporcionales, tales como el uso de residuos, los cuales ayudan a saber si existen problemas con los datos y para saber qué tan bien se ajustan los datos, tal como sucede en el análisis de regresión.

En esta sección se utilizará la base de datos de adictos, esta base está compuesta por 3 variables: prisión, dosis y clínica. La variables prisión y clínica son variables categóricas.

1.7.1. Residuos

Residuos Martingala

Los residuos martingala se utilizan en el análisis de supervivencia principalmente para conocer la forma funcional de las covariables, estas pueden ser dependientes o independientes del tiempo.

Se puede decir que los residuos martingala de manera sencilla son la diferencia entre las fallas observadas y las fallas esperadas de cada individuo en el estudio y se definen como:

$$r_{Mi} = \delta_i - \hat{H}(t_i), \qquad = 1, \ldots, n \tag{1.28}$$

Para realizar el análisis de residuos en R, es necesario primero ajustar un modelo de Cox, después realizar un gráfico de dispersión de cada covariable contra los residuos martingala. Si al graficar los valores de la covariable contra los residuos martingala se obtiene un patrón aleatorio se puede decir que la covariable ingresa lineal al modelo. Los residuos martingala deben calcularse para cada variable en el modelo, sin embargo, si se tienen variables dicotómicas no serán de mucha utilidad.
En el gráfico 1.10 puede verse que los residuos contra los valores de la variable dosis, no están distribuidos de manera aleatoria. Por lo que no puede afirmarse que la covariable dosis tenga una relación lineal con el riesgo.

Residuos Cox-Snell

Los residuos Cox-Snell en general se utilizan para saber si el modelo ajusta de manera correcta a los datos.

Para saber si un modelo está correctamente ajustado al momento de graficar la función estimada de riesgo acumulado contra los residuos del modelo de Cox se deberá

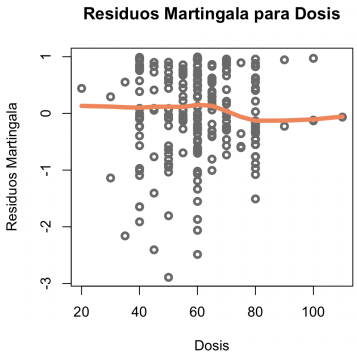

Figura 1.10: Residuos Martingala.

obtener una línea recta con pendiente 1, lo que esto indica es que los residuos se distribuyen exponencial con tasa 1, la demostración puede encontrarse en [Collet, 2015]. El estimador Nelson-Aalen se utiliza para estimar la función base de riesgo acumulado,

$$r_{cs_i} = \exp(\hat{\beta}' X_i)\hat{H}_0(t_i)$$

Para calcular los residuos Cox-Snell en R es necesario restar los residuos martingala y el estatus de cada individuo, denotado con $\delta_i = \{1, 0\}$.

$$Cox - Snell = \delta_i - r_{Mi}$$

En la figura 1.11 puede verse que el modelo ajusta correctamente los datos pues los residuos contra la función de riesgo acumulada siguen la línea recta de pendiente 1. A continuación se muestra el código para calcular los residuos Cox-Snell en R.

```
ajuste <- coxph(Surv(time,status) ~ clinic+prision+ dosis,
method='breslow', data=adictos)
coxsnellres <- adictos$status-ajuste$residuals
ajuste2=survfit(Surv(coxsnellres,adictos$status)~1)
Htilde=cumsum(ajuste2$n.event/ajuste2$n.risk)
plot(ajuste2$time,Htilde,type='p',col='black',
lwd=2,xlab='Residuos Cox-Snell',
    ylab='Función estimada de riesgo acumulado')
abline(0,1,col='red',lty=2,lwd=3)
```

Residuos Schoenfeld

A diferencia de los residuos martingala y de la devianza (los cuales se presentarán mas adelante), la idea principal de los residuos Schoenfeld es brindar un residuo para cada variable e individuo en el modelo para identificar si la covariable en el modelo cumple

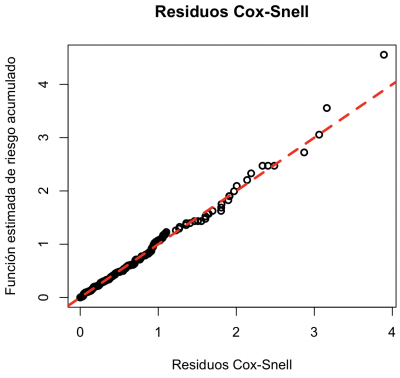

Figura 1.11: Residuos Cox-Snell.

el supuesto de riesgos proporcionales. Estos residuos representan la diferencia entre la covariable observada y el promedio sobre el conjunto en riesgo a cada tiempo de falla, representado en (1.29) como $\hat{a}_{ji}$. El i-ésimo residuo Schoenfeld para la covariable X_j en el modelo está dado por:

$$\begin{aligned} r_{sji} &= \delta_i\{x_{ji} - \hat{a}_{ji}\} \\ \hat{a}_{ji} &= \frac{\sum_{l\in R(t_i)} x_{jl}\exp(\hat{\beta}' x_l)}{\sum_{l\in R(t_i)}\exp(\hat{\beta}' x_l)} \end{aligned} \tag{1.29}$$

donde x_{ji} es el valor de la j-ésima variable explicativa del modelo $j = 1, 2, ...p$, para el i-ésimo individuo en el estudio.

Al hacer el gráfico de los residuos obtenidos la interpretación se vuelve un poco complicada, por lo que generalmente se reescalan los residuos para poder interpretarlos, esto debido a que son parecidos a los obtenidos de estimar vía mínimos cuadrados. En la ecuación (1.30) se presenta la forma de calcular estos residuos, donde d representa el número de fallas y $var(\hat{\beta})$ la matriz $p \times p$ de varianzas y covarianzas de los parámetros estimados obtenida de estimar el modelo de riesgos proporcionales.

$$r^*_{Si} = d \times var(\hat{\beta}) \times r_{Si} \tag{1.30}$$

En la figura 1.12 se observan los residuos Schoenfeld para cada una de las 3 covariables en el modelo de la base adictos. Puede observarse que la covariable prisión y dosis cumplen con el supuesto de riesgos proporcionales, pues la línea ajustada a los datos no depende de estos, tal como lo hace la ajustada a la covariable clínica.

Residuos Score o de puntajes

Los residuos *score* son utilizados para verificar la influencia individual sobre el modelo, además como se verá más adelante, estos residuos sirven para obtener una varianza

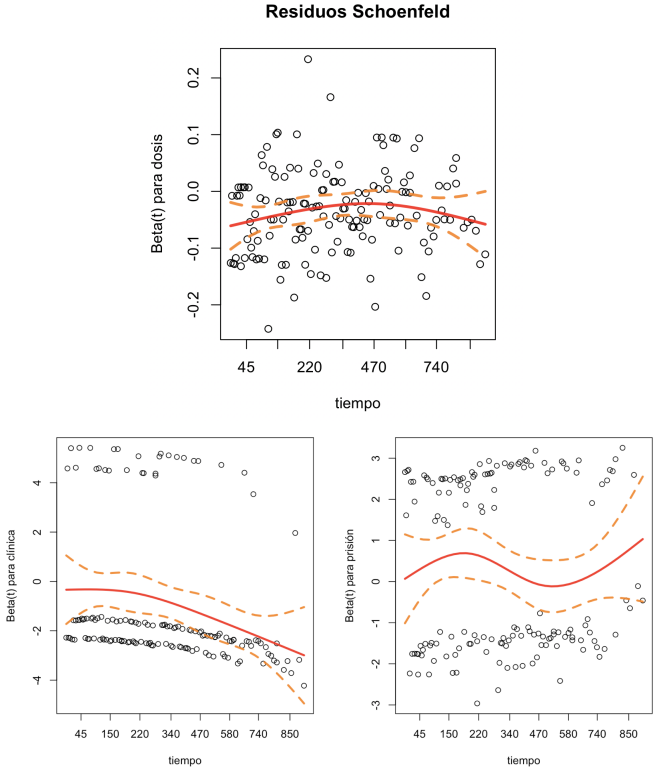

Figura 1.12: Residuos Schoenfeld.

robusta.

Estos residuos se obtienen a través de derivar la función de log-verosimilitud parcial y obtener las derivadas parciales de cada β_j, $j = 1, \ldots, p$,

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta_1} &= 0 \\ &\vdots \\ \frac{\partial \ln L}{\partial \beta_p} &= 0 \end{aligned}$$

Con esto se obtiene un vector de residuos denotado como:

$$r_i' = (r_1, r_2, \ldots, r_p)$$

Residuos de devianza

Los residuos martingala, mencionados previamente, no cumplen con una distribución simétrica alrededor del cero, para intentar solucionar esto se desarrollaron los residuos de devianza.

$$r_{Di} = sgn(r_{Mi})[-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}]^{1/2} \tag{1.31}$$

donde r_{Mi} se refiere a los residuos martingala y sgn a la función signo. Estos residuos

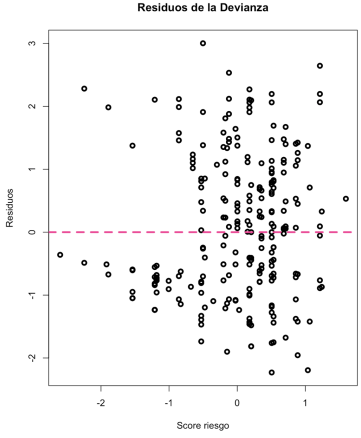

Figura 1.13: Residuos de la Devianza.

son una transformación de los residuos martingala, pero hay que tener en cuenta que los residuos de la devianza a pesar de que se distribuyen alrededor de cero, no necesariamente suman cero. Sin embargo, estos residuos son los más parecidos a los de la regresión lineal (media 0 y varianza 1), además al graficar los valores obtenidos contra los valores de las covariables para saber si las observaciones tienen varianza constante, media cero o si hay observaciones con residuos grandes (2 o 3 en valor absoluto). En el caso del modelo de adictos, en la figura 1.13 puede verse que el modelo tiene varianza constante y no se tienen residuos mas grandes que 3.

Observaciones influyentes

Para poder llevar a cabo un análisis de los datos adecuado es necesario identificar si en el modelo se tiene observaciones influyentes, estas observaciones se llaman de esta forma debido a que si alguna de ellas es eliminada del modelo, las estimaciones en el modelo cambian notablemente.

La manera más costosa computacionalmente de encontrar las observaciones influyentes es ajustar el modelo completo y seguir haciendo ajustes elmininando la j-ésima

observación ($\hat{\beta} - \hat{\beta}_j$). Si se estuviera modelando miles de datos, llevar a cabo este método llevaría mucho tiempo.

La manera óptima de encontrar a los individuos influyentes es a través del método llamado deltas-betas, utilizando los residuos score.

Sea $r'_{Ui} = (r_{Uji}, r_{Uji}, \cdots, r_{Upi})$ el vector de valores de los residuos score, donde $j = 1, \cdots, p$ el número de covariables y $i = 1, \cdots, n$ el número de observaciones en el modelo. El cambio entre el modelo completo y el modelo sin la i-ésima observación se obtiene del j-ésimo elemento del vector:

$$r'_{Ui} var(\hat{\beta}) \tag{1.32}$$

donde $var(\hat{\beta})$ en la ecuación (1.32) es la matriz de varianzas y covarianzas ajustada del modelo de Cox. El j-ésimo elemento de dicho vector, se conoce como delta-beta y se denota como $\Delta_i \hat{\beta}_j$.

$$\Delta_i \hat{\beta}_j \approx \hat{\beta}_j - \hat{\beta}_j(i)$$

Para ver las observaciones influyentes gráficamente se debe graficar los valores de las $\Delta_i \hat{\beta}_j$ contra los individuos en el modelo. Los gráficos de las observaciones influyentes para el ejemplo de la base de adictos se muestran en la figura 1.14. El individuo id=150 tiene un valor atípico, por lo que podría considerarse como valor influyente, sin embargo la unica manera de saber si realmente lo es, es estimando el modelo con y si la observación.

1.7.2. Métodos Gráficos

Mediante la comparación de las curvas de supervivencia $\log(-\log \hat{S})$ ver si hay posibilidade de que las variables cumplan con el supuesto de riesgos proporcionales.

Para probar que el método de $\log(-\log \hat{S})$ se puede aplicar para verificar los riesgos proporcionales se describirá esta función en términos de la función de riesgos proporcionales de Cox,

$$h(t, X) = h_0(t) \exp\left\{\sum_{i=1}^{p} \beta_i X_i\right\}$$

por la sección anterior se sabe que la función de riesgo y la función de supervivencia están relacionadas. Se define $\psi = \exp\left\{\sum_{i=1}^{p} \beta_i X_i\right\}$.

$$\begin{aligned} S(t, X) &= \exp\left(\int_0^t h(u)du\right) = \exp\left(\int_0^t h_0(u)\psi du\right) \\ &= \exp\left(\psi \int_0^t h_0(u)du\right) = \exp\left(\int_0^t h_0(u)du\right)^{\psi} \\ &= S_0(t)^{\psi} \end{aligned}$$

donde $S_0(t)$ es la función de supervivencia base, asociada a la función de riesgo base. Siguiendo el método de $\log(-\log \hat{S})$ se debe calcular el logarítmo de la función de supervivencia,

$$\log S(t, X) = \psi \log(S_0(t)) \qquad 0 \leq S_0(t) \leq 1 \tag{1.33}$$

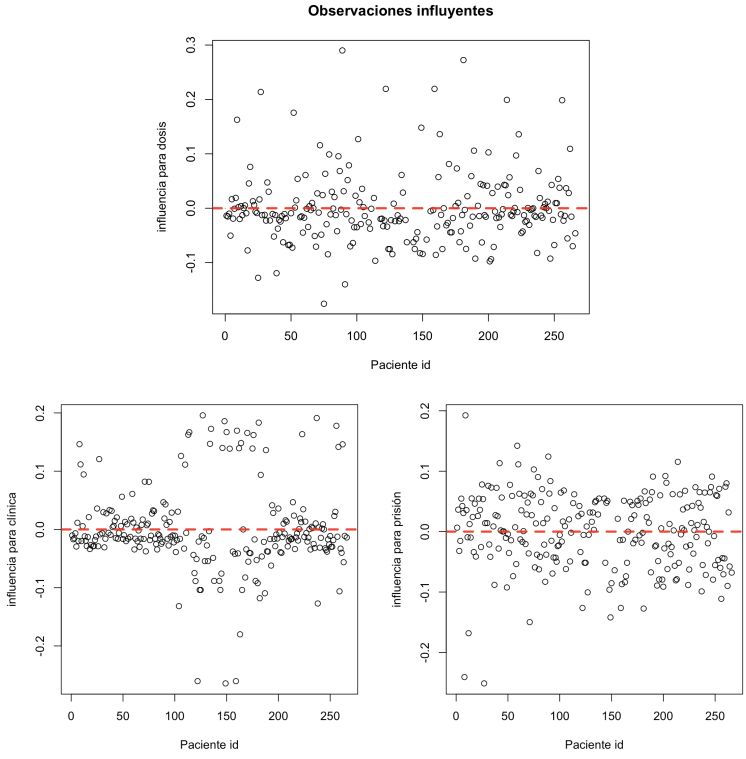

Figura 1.14: Observaciones influyentes en modelo de adictos.

El logaritmo ente 0 y 1 es un número negativo, es por esto que se debe poner el signo menos, para poder calcular el siguiente logaritmo.

$$\begin{aligned} \log(-\log S(t,X)) &= \log\left(\exp\left\{\sum_{i=1}^{p}\beta_i X_i\right\}\log(S_0(t))\right) \\ &= \sum_{i=1}^{p}\beta_i X_i + \log(-\log(S_0(t))) \end{aligned}$$

Se tienen dos sujetos distintos X_1 y X_2, la función de log-log de supervivencia para el sujeto 1, $X_1 = (X_{11}, X_{12}, ..., X_{1p})$ es :

$$\log(-\log S(t,X_1)) = \sum_{i=1}^{p}\beta_i X_{1i} + \log(-\log(S_0(t)))$$

y la del individuo 2, $X_2 = (X_{21}, X_{22}, ..., X_{2p})$ es :

$$\log(-\log S(t, X_2)) = \sum_{i=1}^{p} \beta_i X_{2i} + \log(-\log(S_0(t)))$$

si se restan las funciones se obtiene:

$$\begin{aligned} &\log(-\log S(t, X_1)) - \log(-\log S(t, X_2)) \\ &= \sum_{i=1}^{p} \beta_i X_{1i} + \log(-\log(S_0(t))) - \sum_{i=1}^{p} \beta_i X_{2i} + \log(-\log(S_0(t))) \\ &= \sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i}) \end{aligned}$$

La expresión $\log(-\log S(t))$ para el individuo 1 se puede expresar en términos de la función del individuo 2 más un término independiente del tiempo. Donde la distancia entre las dos curvas será el término independiente. Si dicho término es constante, en general, las gráficas serán paralelas.

$$\log(-\log S(t, X_1)) = \sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i}) + \log(-\log S(t, X_2))$$

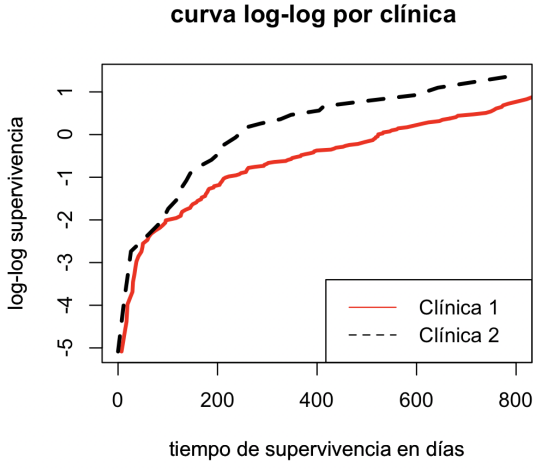

Figura 1.15: Curva log-log de supervivencia.

La sentencia en R para generar el gráfico de la curva log-log para la variable clínica, es la siguiente:

```
plot(clinic1$time, log(-log(clinic1$survival)),
xlab = "tiempo de supervivencia en días",
ylab ="log-log supervivencia",
xlim = c(0,800), col="red", type='l', lty='solid',
main = "curva log-log por clínica ")
par(new=T)
plot(clinic2$time, log(-log(clinic2$survival)),
xlab = "tiempo de supervivencia en días",
ylab ="log-log supervivencia",axes = F, col="black",
type='l', lty='dashed')
legend("bottomright", c("Clínica 1","Clínica 2" ),
lty= c("solid","dashed"), col=c("red","black"))
```

Como puede observarse en la figura 1.15 la función no cumple con el supuesto de riesgos proporcionales para la variable clínica.

1.7.3. Bondad de Ajuste

El método de bondad de ajuste da un estadístico de prueba y un p-value por lo que es más utilizado para probar riesgos proporcionales.

La prueba de bondad de ajuste se basa en los residuos Schoenfeld, cada covariable en el modelo contará con sus residuos Scchoenfeld.

Para llevar a cabo la prueba se deben seguir los siguientes pasos:

- Ajustar un modelo de Cox RP con todas las covariables con lo que se obtendrán los residuos Schoenfeld para cada covariable en el modelo.
- Crear una variable que contenga los tiempos de falla ordenados de forma descendente y numerar.
- Probar la correlación entre los residuos y la variable del punto anterior. $H_0 : \rho = 0$, si se rechaza H_0 se concluye que el supuesto de riesgos proporcionales no se cumple.

Para llevar a cabo el ejemplo en R, se utiliza el resultado de la la función **cox.zph**, si al graficar los residuos la curva ajustada es horizontal entonces se asume que los riesgos son proporcionales pues ésta no depende del tiempo.

En la tabla 1.11 la variable de interés es clínica, dado que el p-value es $0.00119 < 0.05$ se rechaza la hipótesis nula con lo que se afirma que el supuesto de riesgos proporcionales no se cumple para dicha variable. Los residuos Schoenfeld de la variable clínica se encuentran en la figura 1.16, este gráfico surge de graficar cada tiempo de falla de los individuos contra los residuos de Schoenfeld, en este caso la línea ajustada no es totalmente horizontal.

	rho	chisq	p
prisión	-0.04622	0.3215	0.5706
dosis	0.0905	1.0957	0.2952
clínica	-0.2497	10.4952	0.00119
GLOBAL	NA	12.4248	0.0060

Tabla 1.11: prueba de bondad de ajuste

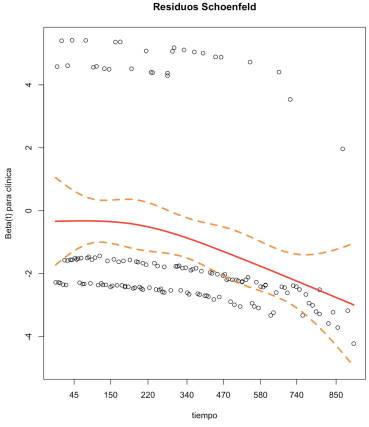

Figura 1.16: Residuos de Schoenfeld.

1.7.4. Covariables Dependientes del Tiempo

Para llevar a cabo el método de las covariables dependientes del tiempo, es necesario extender el modelo de Cox a un modelo con interacción,

$$h(t, X) = h_0(t) \exp\{\beta X + \delta(X \times g(t))\}$$

En el modelo de interacción la variable $g(t)$ se puede escoger de distintas maneras, un ejemplo es:

$$\begin{aligned} g(t) &= t \\ g(t) &= \log(t) \\ g(t) &= \begin{cases} 1 & \text{si } t \geq t_0 \\ 0 & \text{si } t < t_0 \end{cases} \end{aligned}$$

Para probar que se cumple el supuesto de riesgos proporcionales, puede utilizarse la prueba de Wald o la razón de verosimilitudes sobre cada covariable ajustada en el modelo .

En [Kleinbaum and Klein, 2012] puede consultarse un capítulo completo sobre covariables dependientes del tiempo para probar el supuesto de riesgos proporcionales.

Para este ejemplo se utiliza una vez más la base de adictos, el método que se aplica para las interacciones es el de logaritmo del tiempo. Primero se crean las 3 variables dummy que serán parte del modelo.

```
cov1 <- log(adictos$time)* adictos$clinica
cov2 <- log(adictos$time)* adictos$prision
cov3 <- log(adictos$time)* adictos$dosis
ejemplo_cov <- coxph(Surv(time,status) ~clinic+ prision+ dosis +
cov1+cov2+cov3,method='breslow', data=adictos)
```

Tabla 1.12: Ajuste del modelo de Cox para la base de adictos.

	coef	exp(coef)	se(coef)	z	$P(>\|z\|)$
clínica	1.343e+01	6.788e+05	1.543e	8.702	$< 2e-16$ ***
prisión	-1.595	2.030e-01	1.310	-1.217	0.223
dosis	6.175e-01	1.854	5.003e-02	12.343	$< 2e-16$ ***
cov1	-2.559	7.738e-02	2.967e-01	-8.623	$< 2e-16$***
cov2	2.597e-01	1.297	2.293e-01	1.133	0.257
cov3	-1.090e-01	8.968e-01	8.767e-03	-12.430	$< 2e-16$ ***
Rsquare		= 0.935			
Likelihood ratio test		= 650.8	on 6 df,	p=0	
Wald test		= 203	on 6 df,	p=0	
Score (logrank) test		= 464.1	on 6 df,	p=0	

La tabla 1.12 presenta las estimaciones del modelo de Cox con las covariables dummy incluidas en el modelo, las covariables con un p-value menor a 0.05 no cumplen con el supuesto de riesgos proporcionales, tal como se vio con los residuos de Schoenfeld la variable clínica no cumple.

El modelo final, que cumple con los supuestos de riesgos proporcionales es el presentado en la ecuación (1.34).

$$h(t, X) = h_0(t) \exp\{\beta_1 prision + \delta_1 (log(t) * prision)\} \tag{1.34}$$

1.8. Modelo de Cox Estratificado

En ocasiones algunas variables predictoras del modelo de Cox no cumplen con el supuesto de riesgos proporcionales, por lo que no es posible llevar a cabo dicho método y es necesario realizar modificaciones del modelo de riesgos proporcionales.

El nuevo modelo que se plantea es un modelo estratificado de Cox por estrato.

Estratificar un modelo se refiere a crear categorías sobre las variables que no cumplen los supuestos, para así crear estratos que contengan los datos de esas variables y poder llevar a cabo el modelo de Cox.

Esto se puede ver de la siguiente manera, se supone un modelo con k variables que no cumplen con los supuestos de riesgos proporcionales (RP), estas variables se denotan como: $Z_1, Z_2, ..., Z_k$, y p variables que sí lo cumplen: $X_1, X_2, ..., X_p$. Las variables Z_i se categorizarán y combinarán con el fin de obtener los estratos, se denotará como Z^* a la nueva variable que contiene los estratos.
Para ejemplificar el proceso de estratificación, se supone un modelo con 2 covariables que no cumplen con el supuesto de riesgos proporcionales, $k = 2$, estás dos variables son edad y tratamiento, la variable edad se puede categorizar como : joven, adulto y viejo, mientras que tratamiento se divide en 1 y 2. Las combinaciones obtenidas son 6, estas combinaciones serán los estratos y se representarán como k^*. En la table se muestra la estratificación de las variables.

Tabla 1.13: Ejemplo estratos.

	Joven	Adulto	Viejo
Tratamiento 1	1	2	3
Tratamiento 2	4	5	6

Entonces la función de riesgo del modelo estratificado de Cox es la siguiente:

$$h_g(t, X) = h_{0g}(t) \exp\{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\} \qquad g = 1, 2, ..., k^* \tag{1.35}$$

Z^* no está incluida explícitamente en el modelo, pero se toma en cuenta en los estratos. Las funciones de riesgo base h_{0g} pueden ser distintas para cada estrato, lo cual implicará que se obtengan distintas curvas de supervivencia, únicamente la parte exponencial será la misma para todos los niveles.

Para obtener los estimadores de los coeficientes de regresión, se debe maximizar la verosimilitud parcial, que se obtendrá de multiplicar la verosimilitud de cada estrato. Así mismo la función de verosimilitud para cada estrato se obtiene con la función de riesgo por estrato. Esto se ejemplificará a continuación.

Ejemplo

Con el ejemplo del modelo de Cox de fumadores utilizado anteriormente, se ejemplificará el caso estratificado, aumentando la variable asma. Se supone que la variable asma (1 si la tiene, 2 si no la tiene) no cumple con los supuestos de RP, por lo que se va a estratificar.

Tabla 1.14: Ejemplo: modelo estratificado de Cox.

	tiempo	estatus	fumador	asma
Pedro	6	0	1	1
Pablo	8	1	1	2
Juan	5	1	0	1
Luis	2	1	0	2

En la tabla 1.14 se muestra que Pedro y Juan han presentado asma, mientras que Pablo y Juan no.

La estratificación se hará por pasos. Primero se categorizarán los datos según el estatus del asma, con lo que se obtendrán dos grupos, como se muestra en la tabla 1.15.

Tabla 1.15: Datos asma.

Datos asma=1				
	tiempo	estatus	fumador	asma
Pedro	6	0	1	1
Juan	5	1	0	1
Datos asma=2				
	tiempo	estatus	fumador	asma
Pablo	8	1	1	2
Luis	2	1	0	2

Para este ejemplo se tienen 2 estratos, por lo tanto $k^* = 2$. La función de riesgo de cada estrato estará dada como:

$$h_1(t|X) = h_{0_1}(t)\exp\{\beta_1 FUMADOR\} \qquad g = 1 \qquad asma \tag{1.36}$$

$$h_2(t|X) = h_{0_2}(t)\exp\{\beta_1 FUMADOR\} \qquad g = 2 \qquad no\ asma \tag{1.37}$$

Con estas funciones se pueden calcular ahora la función de verosimilitud parcial de cada uno de los estratos. La función de verosimilitud parcial para el estrato 1, donde los sujetos han presentado asma, estará dada únicamente por los datos de Juan al tiempo 5, ya que Pedro está censurado:

$$L_1 = \frac{h_{01}e^{\beta_1 * FUMADOR}}{h_{01}e^{\beta_1 * FUMADOR} + h_{01}e^{\beta_1 * FUMADOR}} = \frac{h_{01}e^{\beta_1 * 0}}{h_{01}e^{\beta_1 * 1} + h_{01}e^{0}} \tag{1.38}$$

La verosimilitud parcial del estrato 2, está formada por ambos sujetos pues ninguno está censurado:

$$L_2 = \frac{h_{02}e^{\beta_1 * FUMADOR}}{h_{02}e^{\beta_1 * FUMADOR} + h_{02}e^{\beta_1 * FUMADOR}} * \frac{h_{02}e^{\beta_1 * FUMADOR}}{h_{02}e^{\beta_1 * FUMADOR}} \tag{1.39}$$

La verosimilitud parcial del modelo estratificado es la siguiente: $L = L_1 * L_2$

$$\begin{aligned}
L(\beta) &= \left[\frac{h_{01}e^{\beta_1(0)}}{h_{01}e^{\beta_1(1)} + h_{01}e^{0}}\right]\left[\frac{h_{02}e^{\beta_1(0)}}{h_{02}e^{\beta_1(0)} + h_{02}e^{\beta_1(1)}} \times \frac{h_{02}e^{\beta_1(1)}}{h_{02}e^{\beta_1(1)}}\right] \\
L(\beta) &= \left[\frac{\cancel{h_{01}}e^{\beta_1(0)}}{\cancel{h_{01}}(e^{\beta_1(1)} + e^{0})}\right]\left[\frac{\cancel{h_{02}}e^{\beta_1(0)}}{\cancel{h_{02}}(e^{\beta_1(0)} + e^{\beta_1(1)})} \times \frac{\cancel{h_{02}}e^{\beta_1(1)}}{\cancel{h_{02}}e^{\beta_1(1)}}\right] \\
L(\beta) &= \left[\frac{e^{0}}{e^{\beta_1} + e^{0}}\right]\left[\frac{e^{0}}{e^{0} + e^{\beta_1}} \times \frac{e^{\beta_1}}{e^{\beta_1}}\right]
\end{aligned} \tag{1.40}$$

Se puede observar que en (1.40) las funciones de riesgo base se cancelan y la verosimilitud parcial queda determinada solamente por el orden de los eventos, tal como en el modelo de Cox de riesgos proporcionales.

1.9. Modelos Paramétricos

El modelo de Cox es uno de los modelos más usados para modelar datos del área de ciencias de la salud, pero no es el único modelo que puede utilizarse, también se tienen modelos de supervivencia conocidos como modelos paramétricos de regresión donde la variable respuesta tal como el tiempo de falla está especificada en términos de parámetros desconocidos. El objetivo en estos modelos es conocer cúal es el efecto que tienen las covariables sobre el modelo.

Los modelos de regresión se dividen principalmente en modelos de riesgos proporcionales, tal como el modelo de Cox, y los modelos de vida acelerada.

Algunos ejemplos de regresión paramétrica conocidos son la regresión lineal, logística y Poisson, donde se asume que la variable respuesta sigue cierta distribución conocida que puede ser la normal, binomial o Poisson, respectivamente, pero los parámetros son

desconocidos. En el caso de los modelos de supervivencia paramétrica se asume que el tiempo de falla sigue cierta distribución conocida: Weibull, exponencial, log-logística, gamma generalizada o log-normal, entre otros

En los modelos paramétricos se conoce la función de densidad y a partir de ésta es posible encontrar la función de riesgo y de supervivencia.

1.9.1. Modelo Exponencial

Trabajar con la función exponencial es sencillo debido a que la función de riesgo es constante $h(t) = \lambda$, debido a la propiedad de pérdida de memoria, esto es, que no importa lo que haya sucedido en el pasado el riesgo será el mismo en cualquier momento.

La función de densidad de la distribución exponencial es $f(t) = S(t)h(t)$ donde $S(t) = \exp(-\lambda t)$ por lo que $f(t) = \lambda \exp(-\lambda t)$.

La función de riesgo del modelo de regresión es:

$$h(t|X) = h_0(t) \exp(\beta' X) \tag{1.41}$$

Si se dice que la función de riesgo base tiene una distribución exponencial se reescribe la ecuación (1.41) como:

$$h(t|X) = \lambda \exp(\beta' X)$$

La función para los modelos paramétricos en R es **survreg**, el problema es que R está diseñado únicamente para ajustar modelos de falla acelerada, por lo que es necesario reparametrizar las variables para obtener los resultados de un modelo de riesgos proporcionales. Sean $\hat{\omega}_0, \hat{\omega}_1, ..., \hat{\omega}_p$ los parámetros estimados por la función.

- $\hat{\lambda} = \exp(-\hat{\omega}_0)$
- $\hat{\beta}_i = -\hat{\omega}_i \qquad \forall i = 1, 2, ..., p$

Para realizar el ejemplo se utilizarán los datos de adictos, utilizada para validar los supuestos del modelo de riesgos proporcionales de Cox. Esta base está compuesta por 3 variables: dosis, prisión y clínica en la que se han atendido los pacientes. La función de riesgo para modelar estos datos es:

$$h(t|X) = \lambda \exp\{\beta_1 clinic + \beta_2 dosis + \beta_3 prision\}$$

La sentencia en R para obtener un ajuste al modelo de tipo exponencial es la siguiente:

```
exp <- survreg(Surv(time,status==1)~
clinica+ prision + dosis, dist="exponential", data=adictos)
summary(exp)
```

Una vez obtenidas las estimaciones de los parámetros es necesario llevar a cabo la transformación antes mencionada. En la tabla 1.16 se muestran las estimaciones del modelo exponencial y en la tabla 1.17 están las estimaciones transformadas

Tabla 1.16: Salida R ajuste modelo exponencial

	Value	Std. Error	z	p
(Intercept)	3.6843	0.4307	8.5539	1.189e-17
adictos$clinica	0.8805	0.2106	4.1807	2.905e-05
adictos$prision	-0.2526	0.1648	-1.5322	1.254e-01
adictos$dosis	0.0289	0.0061	4.7061	2.524e-06

Tabla 1.17: Estimadores modelo exponencial.

	Valor
λ	$\exp(-3.6843)$
β_1	-0.8805
β_2	0.2526
β_3	-0.0289

1.9.2. Modelo Weibull

El modelo Weibull es el modelo paramétrico más utilizado en el análisis de supervivencia. Las funciones asociadas a este modelo se presentan en la tabla 1.18.

Función	
$f(t)$	$\lambda p t^{p-1} \exp(-\lambda t^p)$
$h(t)$	$\lambda p t^{p-1}$
$S(t)$	$\exp(-\lambda t^p)$
$H(t)$	λt^p

Tabla 1.18: Funciones distribución Weibull.

Un modelo de supervivencia sigue la distribución Weibull si cumple que la función de riesgo base se distribuye Weibull con parámetros (p, λ),

$$\begin{aligned} h(t|X) &= h_0(t)\exp(\beta' X) \\ h(t|X) &= \lambda p t^{p-1}\exp(\beta' X) \end{aligned} \tag{1.42}$$

Con la función de riesgo se puede calcular la función de supervivencia.

$$\begin{aligned} S(t|X) &= \exp\left\{-\int_0^t h(u|X)du\right\} \\ S(t|X) &= \exp\left\{-\int_0^t \lambda p u^{p-1}\exp(\beta' X)du\right\} \\ S(t|X) &= \exp\left\{-\exp(\beta' X)\int_0^t \lambda p u^{p-1}du\right\} \\ S(t|X) &= \exp\{-\exp(\beta' X)\lambda t^p\} = \exp\{-\lambda t^p\}^{\exp(\beta' X)} \\ S(t|X) &= S_0(t)^{\exp(\beta' X)} \end{aligned} \tag{1.43}$$

Al implementar el modelo Weibull en R se debe tratar de manera distinta la estimación de parámetros ya que por default R ajusta un modelo de vida acelerada en vez de un modelo de riesgos proporcionales.
La función que toma R en la función **survreg** es la expresada en la ecuación (1.44)

$$S(t|X) = \exp\left\{-\exp\left(\frac{log(t)-\mu-\alpha' X}{\sigma}\right)\right\} \tag{1.44}$$

donde los parámetros de la ecuación (1.43) interpretados en términos de la (1.44) son los siguientes:

- $p = \sigma^{-1}$
- $\lambda = \exp(-\mu/\sigma)$
- $\beta_j = -\alpha_j/\sigma$, $\forall j = 1, ..., p$.

Para el ejemplo del modelo Weibull nuevamente se utiliza la base de adictos.
Si se escribe el modelo como en la expresión (1.44) se obtiene la función de supervivencia:

$$S(t|X) = \exp\left\{-\exp\left(\frac{log(t)-\mu-(\alpha_1 clinica + \alpha_2 dosis + \alpha_3 prision)}{\sigma}\right)\right\}$$

```
mod.weibull <- survreg(Surv(time,status==1)~
clinic+ prision + dosis,
                dist="weibull", data=adictos)
```

En la tabla 1.19 y en 1.20 se muestran las estimaciones de los coeficientes del modelo y la parametrización de estos coeficientes.

Tabla 1.19: Estimadores modelo Weibull.

		Value	Std. Error	z	p
μ	(Intercept)	4.1048	0.32805	12.512	6.3746e-36
α_1	adictos$clinic	0.7090	0.15722	4.509	6.4908e-06
α_2	adictos$prision	-0.2294	0.12078	-1.8997	5.7467e-02
α_3	adictos$dose	0.02442	0.0045	5.321	1.0284e-07
$\log(\sigma)$	Log(scale)	-0.3149	0.06755	-4.661	3.1324e-06

Tabla 1.20: Estimadores modelo Weibull.

parámetro	parametrización	Valor
λ	$\exp(\frac{\mu}{\sigma})$	0.00361
p	σ^{-1}	1.37
β_1	$-\alpha_1/\sigma$	-0.97
β_2	$-\alpha_2/\sigma$	0.31
β_3	$-\alpha_3/\sigma$	-0.0334

1.10. Modelos Frailty

Un modelo frailty incluye un componente aleatorio diseñado para agregar cierta variabilidad no observada en cada individuo que no puede ser tomada en cuenta en los predictores, este componente es conocido como la frailty.

El componente frailty α es un efecto no observado multiplicativo sobre la función de riesgo ($\alpha > 0$) y se asume que sigue una distribución $g(\alpha)$ con media $\mu = 1$ y varianza $\sigma^2 = \theta$ la cual se estima directamente de los datos.

La función de riesgo para un modelo de frailty es $h(t|\alpha) = \alpha h(t)$ y la función de supervivencia $S(t|\alpha) = S(t)^\alpha$.

Cuando un individuo tiene $\alpha > 1$ tendrá mayor riesgo y menor probabilidad de sobrevivir, mientras que $\alpha < 1$ indica lo contrario.

La supervivencia condicional $S(t|\alpha)$ se refiere a la supervivencia individual y la no condicional $S_u(t)$ expresada en (1.45) considera el promedio de la población. Una vez que la distribución de $g(\alpha)$ es escogida la supervivencia no condicional se obtiene integrando sobre la supervivencia condicional por $g(\alpha)$,

$$S_u(t) = \int_0^\infty S(t|\alpha)g(\alpha)d\alpha \tag{1.45}$$

La función de riesgo no condicional, se puede obtener a partir de la ecuación (1.2) y la ecuación (1.6)

$$h_u(t) = -\frac{d[S_u(t)]/dt}{S_u(t)}$$

En R se tienen 3 opciones para elegir la distribución de la *frailty*: Gamma, Gaussiana y t-Student. La varianza θ de la *frailty* es un parámetro estimado por el modelo, si ésta es 0 entonces no hay *frailty*.

1.10.1. Modelo Frailty Compartida

En los modelos *frailty* se le conoce como modelos *frailty* compartida a aquellos modelos que agrupan (crean clusters) a los individuos del estudio con el fin de que compartan la *frailty*. Esto es importante ya que la varianza estimada puede verse como la correlación que existe entre las variables.

El riesgo condicional del sujeto j del cluster k se expresa como $\alpha_k h_{jk}(t)$, donde $h_{jk}(t)$ depende de las covariables X_{jk}. La frailty α_k únicamente depende del cluster y no del individuo pues es compartida.

$$h_{jk}(t|\alpha_k) = \alpha_k h_{jk}(t) \qquad j = 1, 2, ..., n_k$$

Adicionalmente se puede agregar la frailty compartida a un modelo de Cox, para conocer la correlación entre los individuos.

$$h_{jk}(t|\alpha_k) = \alpha_k h_0(t) \exp(\beta X_{jk}) \qquad j = 1, 2, ..., n_k$$

Cuando se incluye la *frailty* en los modelos de Cox, en ocasiones el supuesto de riesgos proporcionales puede violarse por lo que se debe ser cuidadoso al interpretar los coeficientes estimados, únicamente se podrá utilizar este método para calcular la razón de riesgo condicionada para la misma *frailty*.

1.11. Procesos de conteo

En el modelo de Cox, la supervivencia está expresada como una función de tres indicadores del tiempo de vida, éstas son: t_i, δ_i, x_i, donde t_i representa el tiempo de observación de la variable, δ_i es una indicadora que puede tomar dos valores, 1 en caso de que ocurra la falla y 0 en caso de censura, x_i representa el vector de covariables. En un proceso de conteo se reemplazan las variables por las funciones del tipo $\{N_i(t), Y_i(t), Z_i(t)\}$, las cuales se describen a continuación.

El proceso $N_i(t)$ se refiere al número de eventos observados en el tiempo $(0, t]$ para el individuo $i = 1, 2, ..., n$. $N_i(t) = I(T_i < t, \delta_i = 1)$ donde $\delta_i = \{0/1\}$ (0 si es censurado, 1 si ocurre la falla). $N(s, t) = N(t) - N(s)$ es el número de eventos ocurridos en el intervalo $(s, t]$ y $N(t) = N(0, t)$ con $t > 0$. La función $N(t)$ es una función escalonada continua

por la derecha con saltos de una unidad.

El proceso $Y_i(t)$ denota si el individuo i está en riesgo justo antes de t. La variable $Y_i(t) = I\{\tilde{T}_i \leq t\}$ es un proceso continuo por la izquierda que depende del tiempo mínimo que haya sucedido entre el tiempo de falla (T_i) y el tiempo de censura (C_i), $\tilde{T}_i = \text{mín}\{T_i, C_i\}$ y $Y(t) = \sum_{i=1}^{n} Y_i(t)$. La figura 1.17 representa gráficamente $\tilde{T}_i$, donde $*$ representa la falla y $\circ$ la censura a lo largo del tiempo t.

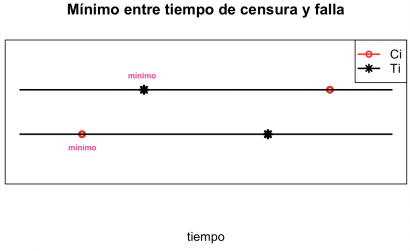

Figura 1.17: Mínimo entre tiempo de falla y tiempo de censura.

Por último $Z(t)$ es el vector de covariables, equivalente a X en el modelo de Cox. En los procesos de conteo usualmente se escriben dependientes del tiempo.

Los procesos de conteo están basados en el histórico de la supervivencia, censura y covariables de un evento, por lo que las variables están referidas a una filtración: $\{\mathcal{F}_t : t \leq 0\}$. La filtración es una sub-σ-álgebra de la σ-álgebra $\mathcal{F}$ donde $\mathcal{F}_t = \sigma\{N_i(s), Z_i(s), Y_i(s), i = 1, ..., n : 0 \leq s \leq t\}$, $t > 0$.

El límite por la izquierda se denota como $\mathcal{F}_{t^-}$ y se refiere a la σ-álgebra generada por los procesos estocásticos $\{N(s), Z(s), Y(s)\}$ en $[0, t)$. Además para toda $s \leq t$ implica que $\mathcal{F}_s \subset \mathcal{F}_t$. A la filtración hasta el tiempo justo antes de t se le conoce también como el histórico del proceso y se denota como $H(t)$.

Sea $dN_i(t)$ el cambio infinitesimal en el proceso entre $(t, t+\Delta t)$. Y además se tiene que la filtración $\mathcal{F}_{t^-}$ contiene toda la información del proceso en $[0, t)$. Entonces la esperanza de que en $dN_i(t)$ se tenga un evento dado el histórico del proceso es la siguiente:

$$E[dN_i(t)|\mathcal{F}_{t^-}] = 1 \times P[dN_i(t)|\mathcal{F}_{t^-}] + 0 \times P[dN_i(t)|\mathcal{F}_{t^-}] = P[dN_i(t)|\mathcal{F}_{t^-}] \qquad (1.46)$$

Por otra parte, lo relevante de la historia de un sujeto es únicamente su estado al tiempo t^-, es decir si está en riesgo o no, dicha información está contenida en la variable $Y_i(t)$,

por lo tanto la ecuación (1.46) se reescribe sustituyendo la filtración por el estado de la variable.

$$P[dN_i(t) = 1|\mathcal{F}_{t^-}] = P[dN_i(t) = 1|Y_i(t)]$$

Si $Y_i(t) = 0$ querrá decir que el sujeto ya no está en riesgo pues ya no está en observación o bien ha presentado alguna falla, por lo tanto $dN_i(t) = 0$ con probabilidad 1, pues no hay manera de que ocurra algún evento, por lo que solo se toma en cuenta cuando $Y_i(t) = 1$. El hecho de que $Y_i(t) = 1$ implica que el individuo está en riesgo, es decir que puede presentar un evento, por lo que en el tiempo t el individuo puede presentar una censura o una falla, es por esto que en (1.47) se puede sustituir a $Y_i(t)$ por $t \leq T_i, t \leq C_i$

$$P[dN_i(t) = 1|Y_i(t) = 1] = P[t \leq T_i < t + dt|t \leq T_i, t \leq C_i] \tag{1.47}$$

Si el individuo i-ésimo se censuro $C_i = 1$ no puede ocurrir una falla en $dN_i(t)$ por lo tanto se puede expresar la ecuación (1.47) para el caso de no censura como :

$$P[t \leq T_i < t + dt|t \leq T_i, t \leq C_i] = P[t \leq T_i < t + dt|t \leq T_i] = \lambda(t)dt \tag{1.48}$$

por la definición de función de riesgo (1.3). Entonces, se puede combinar la posibilidad de que el sujeto esté o no en riesgo con la siguiente ecuación, con lo que se define la esperanza de que ocurra un evento:

$$P[dN_i(t) = 1|Y_i(t)] = Y_i(t)\lambda(t)dt$$

La función de riesgo para el individuo i al tiempo t, dada la función λ_0 se puede expresar como el proceso de conteo:

$$\lambda_i(t) = \lambda_0(t) \exp\{Z_i'(t)\beta\}$$

donde, como se mencionó en el modelo de Cox, β es el vector de coeficientes de las covariables. Por lo tanto, la función de verosimilitud parcial para n ternas independientes $\{N_i, Z_i, Y_i\}, i = 1, ..., n$ está dada por:

$$\begin{aligned} Lp(\beta) &= \prod_{i=1}^{n} \prod_{t \geq 0} \left[\frac{Y_i \exp\{Z_i'(t)\beta\}}{\sum_{l \in R(t_{(i)})} Y_l \exp\{Z_l'(t)\beta\}} \right]^{dNi(t)} \\ \log Lp(\beta) &= \sum_{i=1}^{n} \int_0^{\infty} dNi(t) \left[Z_i'(t)\beta - \log \left\{ \sum_{l \in R(t_{(i)})} Y_l \exp\{Z_l'(t)\beta\} \right\} \right] \end{aligned} \tag{1.49}$$

para encontrar el estimador $\hat{\beta}$ se debe igualar a cero la derivada parcial con respecto a β de (1.49),

$$\frac{\partial \log Lp(\beta)}{\partial \beta} = \sum_{i=1}^{n} \int_0^{\infty} [Z_i(s) - \bar{x}(\beta, s)] \, dN_i(s)$$

donde $\bar{x}(\beta, s)$ es una media ponderada de las covariables de los individuos que están en riesgo al tiempo s,

$$\bar{x}(\beta, s) = \frac{\sum_{i=1}^{n} Y_i(s) \exp\{Z_i'(s)\beta\} Z_i(s)}{\sum_{i=1}^{n} Y_i(s) \exp\{Z_i'(s)\beta\}} \tag{1.50}$$

Para eventos recurrentes se tienen dos tipos de procesos de conteo, el proceso Poisson y los procesos de renovación.

Un proceso Poisson describe situaciones donde los eventos ocurren de manera aleatoria e independiente, y son comúnmente utilizados en eventos que son influenciados o provocados por factores externos aleatorios, mientras que los procesos de renovación describen tiempos de espera entre un evento y otro independientes, y son utilizados para explicar escenarios en donde los eventos dependen de ciclos internos de un sistema o individuo.

Capítulo 2

Supervivencia con Eventos Recurrentes

Al realizar un análisis de supervivencia el principal interés es conocer los tiempos de falla de los individuos dentro de un estudio, sin embargo en ocasiones no solamente interesa la primera falla sino saber si el individuo presenta más fallas a través del tiempo, estas fallas no deben ser precisamente todas del mismo tipo, también es de interés conocer los tiempos en los que ocurrieron las repetidas recaídas. Los eventos que suceden mas de una vez son conocidos como eventos recurrentes.

Las mayores investigaciones del tema se han desarrollado con el fin de modelar estudios médicos, donde no solo es importante conocer cuando un paciente presentó el primer evento o si el paciente murió. Por ejemplo, supóngase que se está realizando un estudio sobre la disminución de ataques al corazón después de aplicar cierto tratamiento, donde no solo importa si el paciente vuelve a presentar un ataque sino cuántos ataques presenta durante el tiempo de observación. Este deseo por conocer las recurrencias no se ha quedado unicamente en el área de la salud, sino que se ha extendido a otras áreas, por ejemplo en las aseguradoras interesa conocer cuántas veces entra una reclamación a la compañía o los tiempos en los que llegan éstas.

En este capítulo se mostrarán los modelos que se han ido desarrollando para modelar los eventos recurrentes. El desarrollo de estos modelos surge como una extensión del modelo de Cox, tomando en cuenta que no solamente ocurre un evento. Para esto es necesario utilizar los procesos de conteo, para medir de forma adecuada la recurrencia de los eventos.

2.1. Modelos Semiparamétricos

Para modelar la supervivencia con eventos recurrentes, principalmente se utilizan modelos semiparamétricos. Los modelos semiparamétricos que se han desarrollado y es-

tudiado para el análisis de datos recurrentes son:

- Modelo Andersen-Gill (AG) (1982)
- Modelos Prentice-William-Peterson (PWP) (1981)
- Modelo Wei-Lin-Weissfeld (WLW) (1989)

El modelo Andersen-Gill es el más utilizado en la modelación de múltiples eventos, además es el más sencillo de los 3 modelos. La diferencia entre los tres modelos radica en cómo se quieran tratar las recurrencias.

Si se supone que todos los eventos son idénticos, sin importar el orden en el que ocurrieron, entonces lo mejor es utilizar un modelo de Andersen-Gill. Si lo importante es el orden o categorizar las fallas, entonces se debe recurrir a un modelo de Prentice, William y Peterson, conocido también como modelo condicional. En cambio si se quiere tratar cada evento como distinto conviene utilizar un modelo marginal (Wei-Lin-Weissfeld).

La principal diferencia entre usar el modelo de Cox para eventos recurrentes y no recurrentes es la forma en que éstos se manejan en la función de verosimilitud, ya que en eventos recurrentes los elementos solamente se eliminan del conjunto en riesgo si presentan censuras. Sin embargo, tanto en recurrentes y no recurrentes los individuos se tratan como independientes a pesar de que en recurrencia un individuo puede estar relacionado a varias fallas.

Cada uno de los 4 modelos que se presentarán a lo largo de este capítulo responde a una pregunta de investigación distinta. El modelo de AG es una generalización del modelo de Cox y el resultado de interés se encuentra en el tiempo desde que un individuo se expuso a cierto tratamiento y el momento en que ocurre el evento. Utilizando una función de riesgo base igual para todos los individuos y estimando un parámetro global para todos los factores de interés. Este modelo asume que las fallas en los individuos son independientes y en caso de que este supuesto no se cumpla, se utiliza una matriz de covarianzas sandwich robusta para los estimadores, con lo que se obtendrán errores estándar robustos. Los modelos PWP tienen dos enfoques distintos en los tiempos de recurrencia, el modelo PWP1 ajusta un modelo sobre el tiempo total, tal como en el modelo AG, este modelo evalúa el efecto de las covariables para el k-ésimo evento desde que el individuo entró al estudio. El modelo PWP2 toma en cuenta el tiempo entre eventos y evalúa el efecto de las covariables entre eventos, además estos modelos pueden utilizarse cuando el interés está en predecir un próximo evento. Finalmente el modelo WLW, también conocido como modelo marginal se puede interpretar en términos del promedio de eventos si no hay covariables dependientes del tiempo. Este modelo es apropiado cuando la estructura de dependencia no es de interés.

2.1.1. Varianza robusta

Asumir que las múltiples observaciones en un individuo son independientes, puede ocasionar que los estimadores no sean los mejores ya que las observaciones provenientes

de un mismo individuo pueden, en ocasiones, estar correlacionadas. Para solucionar esto, se debe ajustar las varianzas estimadas de los coeficientes de regresión obtenidas, cabe señalar que los estimadores $\hat{\beta}$ no se ajustan, solo sus varianzas $\widehat{Var(\hat{\beta})}$. El método que se utiliza para estimar las varianzas es el de estimación robusta, mediante un estimador Jacknife agrupado.

La idea del estimador Jacknife agrupado es ir eliminando a un individuo de la muestra cada vez y reestimar el modelo que resulta, la misma idea de los residuos dfbeta para encontrar individuos influyentes en el estudio.

Obtener estimadores robustos sobre la varianza permite calcular intervalos de confianza, así como realizar pruebas de hipótesis, en los modelos expresados en este capítulo. En el modelo de Cox la estimación de la varianza $(\hat{\beta})$ se obtiene con la inversa de la matriz de información observada denotada como $I^{-1}(\hat{\beta})$ de tamaño $p \times p$

La matriz de varianzas y covarianzas robustas se encuentra de la siguiente manera:

$$I'(\hat{\beta})A(\hat{\beta})I^{-1}(\hat{\beta})$$

donde $A(\hat{\beta}) = U'(\hat{\beta})U(\hat{\beta})$ es un factor de corrección de tamaño nxp, este factor se obtiene de multiplicar por ella misma la matriz de coeficientes score.

Finalmente el estimador sandwich para obtener la varianza robusta donde $D(\hat{\beta}) = U(\hat{\beta})I^{-1}(\hat{\beta})$

$$\begin{aligned} \widehat{Var}(\hat{\beta}) &= I^{-1}(\hat{\beta})U'(\hat{\beta})U(\hat{\beta})I^{-1}(\hat{\beta}) \\ \widehat{Var}(\hat{\beta}) &= D'(\hat{\beta})D(\hat{\beta}) \end{aligned}$$

2.1.2. Modelo Andersen-Gill

En 1982 Andersen-Gill propusieron un modelo semiparamétrico de riesgos proporcionales para eventos recurrentes, es decir, que este modelo es una generalización del modelo de Cox.

En este modelo el individuo estará en riesgo durante todo el estudio, a menos que sea censurado.

La función de verosimilitud parcial para este modelo está dada por el producto de la verosimilitud de cada tiempo L_i, tal como en el modelo de riesgos proporcionales de Cox.

$$L = \prod_{i=1}^{n} \prod_{t>0} \left\{ \frac{Y_i(t)\exp\{\beta' X_i\}}{\sum_{j=1}^{n} Y_j(t)\exp\{\beta' X_i\}} \right\} \tag{2.1}$$

La función de riesgo para el i-ésimo individuo está dada por:

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{Z_i'(t)\beta\} \tag{2.2}$$

donde $Y_i(t)$ indica si el individuo i-ésimo está en riesgo o no. En caso de que se presente una falla, el valor de la variable $Y_i(t) = 1$, pero si ocurre alguna censura este cambiará a $Y_i(t) = 0$ pues ya el sujeto dejará de estar en riesgo.

Ejemplo

Supóngase que se está llevando un estudio sobre 85 pacientes que tuvieron algún tumor por cáncer de vejiga. El objetivo de este estudio es modelar la tasa de intensidad con la que los tumores vuelven a aparecer después de que son eliminados. Estos datos se pueden encontrar en la paqueteria *survival* del software R. La base de datos se llama *bladder* y está dividida en 3 bases. La que se utilizará en este ejemplo se llama *bladder1*.

Para llevar a cabo un análisis de datos recurrentes se debe tomar en cuenta que el manejo de los datos es distinto al del análisis al primer evento. En los eventos recurrentes se tienen intervalos del tipo $(t_{j-1}, t_j]$ donde t_{j-1} representa el tiempo en el que se empieza a observar al individuo y t_j el momento en el que ocurre la falla.

Con fines de ilustrar el diseño de los datos, a continuación se mostrarán los primeros 26 sujetos en el estudio. Los individuos que están dentro del estudio pueden tener intervalos de tiempo distintos.

Las covariables que se utilizan para explicar el modelo pueden o no depender del tiempo, una variable que no depende del tiempo se refiere a que no cambiará a lo largo del estudio, por ejemplo si la variable X_1 es el género, entonces ésta será 0 si es hombre, 1 si es mujer, sin importar cuanto tiempo dure el estudio. Pero si por el contrario X_2 es la medida de estrés diaria en un individuo, entonces en este caso los valores cambiarán en cada intervalo, pues no es una condición fija a través del tiempo. Los datos necesarios para poder llevar a cabo el ejemplo son los siguientes:

- N - número de individuos.
- r_i - intervalos de tiempo para el sujeto i.
- δ_{ij} - (0 o 1) estatus del evento para el individuo i en el intervalo j.
- t_{ij0}- tiempo de inicio para el sujeto i en el intervalo j.
- t_{ij1} - tiempo de paro para el sujeto i en el intervalo j.
- X_{ijk} - valor del k-ésimo predictor para el sujeto i en el intervalo j.

Donde $i = 1, ..., N; j = 1, ..., n_i; k = 1, 2, ..., p$.

Los datos de los primeros 26 individuos en el estudio representados de la forma mencionada anteriormente se encuentran en la tabla 2.1. Las variables pueden tomar dos valores, $tx = 1$ si se aplicó el tratamiento de *thioterapia* y 0 si fue placebo. Las variables *num* y *size* se refieren al número y tamaño inicial de tumores, respectivamente, para este ejemplo la variable de interés es el tipo de tratamiento.

Si se observa al sujeto $id = 1$, se puede ver que no presentó ninguna falla pues después de 1 mes de observación salió del estudio. A este sujeto se le aplicó el tratamiento de *thioterapia*. Por otra parte, el sujeto $id = 8$ al mes 5 presentó la reaparición de tumor, el cual se eliminó y el sujeto continuó en el conjunto en riesgo para que al mes 18 saliera del estudio.

Tabla 2.1: Datos de los primeros 26 individuos en el estudio.

id	**int**	**start**	**stop**	**event**	**rx**	**num**	**size**
N	r_i	t_{ij0}	t_{ij1}	δ_{ij}	X_{ij1}	X_{ij2}	X_{ij3}
1	1	0	1	0	1	1	3
2	1	0	4	0	1	2	1
3	1	0	7	0	1	1	1
4	1	0	10	0	1	5	1
5	1	0	6	1	1	4	1
5	2	6	10	0	1	4	1
6	1	0	14	0	1	1	1
7	1	0	18	0	1	1	1
8	1	0	5	1	1	1	3
8	2	5	18	0	1	1	3
9	1	0	12	1	1	1	1
9	2	12	16	1	1	1	1
9	3	16	18	0	1	1	1
10	1	0	23	0	1	3	3
11	1	0	10	1	1	1	3
11	2	10	15	1	1	1	3
11	3	15	23	0	1	1	3
12	1	0	3	1	1	1	1
12	2	3	16	1	1	1	1
12	3	16	23	1	1	1	1
13	1	0	3	1	1	3	1
13	2	3	9	1	1	3	1
13	3	9	21	1	1	3	1
13	4	21	23	0	1	3	1
14	1	0	7	1	1	2	3
14	2	7	10	1	1	2	3
14	3	10	16	1	1	2	3
14	4	16	24	1	1	2	3
15	1	0	3	1	1	1	1
15	2	3	15	1	1	1	1
15	3	15	25	1	1	1	1
16	1	0	26	0	1	1	2
17	1	0	1	1	1	8	1
17	2	1	26	0	1	8	1
18	1	0	2	1	1	1	4
18	2	2	26	1	1	1	4
19	1	0	25	1	1	1	2
19	2	25	28	0	1	1	2
20	1	0	29	0	1	1	4
21	1	0	29	0	1	1	2

Tabla 2.1: Datos de los primeros 26 individuos en el estudio.

id	**int**	**start**	**stop**	**event**	**rx**	**num**	**size**
N	r_i	t_{ij0}	t_{ij1}	δ_{ij}	X_{ij1}	X_{ij2}	X_{ij3}
22	1	0	29	0	1	4	1
23	1	0	28	1	1	1	6
23	2	28	30	1	1	1	6
24	1	0	2	1	1	1	5
24	2	2	17	1	1	1	5
24	3	17	22	1	1	1	5
24	4	22	30	0	1	1	5
25	1	0	3	1	1	2	1
25	2	3	6	1	1	2	1
25	3	6	8	1	1	2	1
25	4	8	12	1	1	2	1
26	1	0	12	1	1	1	3
26	2	12	15	1	1	1	3
26	3	15	24	1	1	1	3
26	4	24	31	0	1	1	3

En este ejemplo el número de individuos en riesgo es 26, entonces $n_{(0)} = 26$ en $t_{(0)}$.

En la figura 2.1 se muestran las múltiples ocurrencias que tuvo cada individuo durante 31 meses de observación. Las fallas, es decir, la reaparición de tumor están representadas con $\times$, y los $\circ$ muestran las censuras. Se puede observar que los sujetos con $id = 1, 2, 3, 4, 6, 7, 10, 16, 20, 21, 22$ presentaron solamente un evento, censura o falla. El máximo de eventos presentados en este ejemplo es 4, los individuos que experimentaron 4 eventos son los etiquetados con id= 13,14,24,25,26.

En la tabla 2.2, se muestran los tiempos de falla del ejemplo anterior ordenados.

Cuando se están observando eventos recurrentes no es posible sacar a los individuos del experimento una vez que presentaron una falla, a diferencia del modelo de riesgos proporcionales de Cox. Para el caso de recurrencias los eventos son todos tratados como independientes aun cuando varios provengan del mismo individuo. Es por eso que el número de individuos en riesgo disminuirá únicamente cuando se presente alguna censura. A partir del primer mes se empezó a tener eventos, el individuo 1 salió del conjunto en riesgo al tiempo 1 pues presentó una censura lo que ocasionó que el conjunto se redujera de 26 a 25, en el mismo intervalo de tiempo falló el individuo 17, pero como sigue en riesgo no se debe eliminar del conjunto, al tiempo 26 el individuo 17 presenta un nuevo evento pero esta vez es una censura (ver tabla 2.1) por lo que se eliminará del estudio y sí disminuirá el conjunto en riesgo.

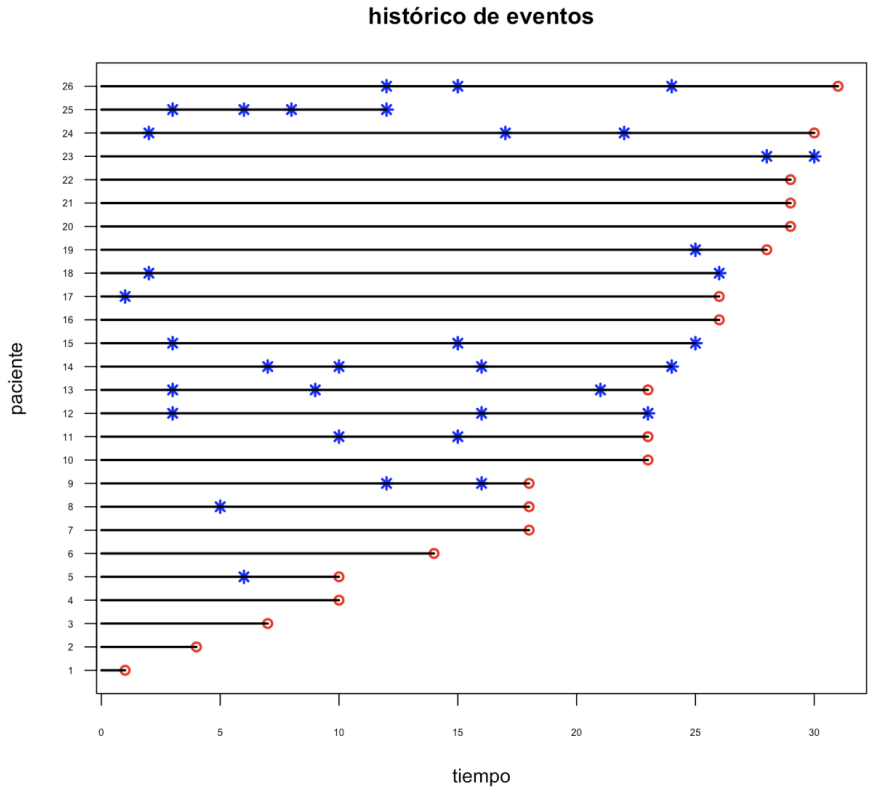

Figura 2.1: Histórico de fallas.

La verosimilitud parcial del conjunto de datos, se expresa como el producto de la verosimilitud de cada individuo. Por lo tanto la verosimilitud parcial de los 26 individuos mostrados en la figura 2.1 es la siguiente:

$$L = L_1 * L_2 * L_3 * ... * L_{26}$$

La verosimilitud del j-ésimo individuo muestra la probabilidad condicional de fallar en el tiempo t_j dado que el individuo sigue en el conjunto en riesgo al tiempo t_j,

$$L_f = \frac{\exp(\beta_1 rx_{(f)} + \beta_2 num_{(f)} + \beta_3 size_{(f)})}{\sum_{s \in R(t_{(f)})} \exp(\beta_1 rx_{s(f)} + \beta_2 num_{s(f)} + \beta_3 size_{s(f)})} \qquad f = 1, \ldots, 26 \qquad (2.3)$$

Tabla 2.2: Tiempos de falla ordenados.

intervalo $[t_j, t_{j+1})$	**tiempo** t_j	**no. riesgo** n_j	**no. fallas** m_j	**censuras** $[t_j, t_{j+1})$	**id** $[t_j, t_{j+1})$
$[0, 1)$	0	26	0	0	0
$[1, 2)$	1	26	1	1	1,17
$[2, 3)$	2	25	2	0	18,24
$[3, 4)$	3	25	4	0	12,13,15,25
$[4, 5)$	4	25	0	1	2
$[5, 6)$	5	24	1	0	8
$[6, 7)$	6	24	2	0	5,25
$[7, 8)$	7	24	1	1	3,14
$[8, 9)$	8	23	1	0	25
$[9, 10)$	9	23	1	0	13
$[10, 11)$	10	23	2	2	4,5,11,14
$[12, 13)$	12	21	3	0	9,25,26
$[14, 15)$	14	21	0	1	6
$[15, 16)$	15	20	3	0	11,15,26
$[16, 17)$	16	20	3	0	9,12,14
$[17, 18)$	17	20	1	0	24
$[18, 19)$	18	20	0	3	7,8,9
$[21, 22)$	21	17	1	0	13
$[22, 23)$	22	17	1	0	24
$[23, 24)$	23	17	1	3	10,11,12,13
$[24, 25)$	24	14	2	0	14,26
$[25, 26)$	25	14	2	0	15,19
$[26, 27)$	26	14	1	2	16,17,18
$[28, 29)$	28	12	1	1	19,23
$[29, 30)$	29	11	0	3	20,21,22
$[30, 31)$	30	8	1	1	23,24
$[31, 32)$	31	7	0	1	26

En la ecuación (2.3) las expresiones $rx_{(f)}, num_{(f)}, size_{(f)}$ corresponden a los valores de las variables *num*,*size*,*rx* del individuo que falla al tiempo $t_{(f)}$. Mientras que los términos $rx_{s(f)}, num_{s(f)}, size_{s(f)}$ representan los valores de las variables del modelo para el sujeto s en el conjunto en riesgo $R_{(t_{(f)})}$, este conjunto contiene a todos los sujetos que aún están en riesgo de falla al tiempo $t_{(f)}$.

Por ejemplo, el sujeto $id = 13$ falló por segunda vez al tiempo $t = 21$, este tiempo corresponde a el tiempo ordenado $f = 18$, en la figura 2.2, se puede observar que al

tiempo 21 el conjunto en riesgo es $n_f = 17$.
$R(t_{(18)} = 21) = \{id = 7, 8, 9, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26\}$. La verosimilitud del tiempo ordenado $t_{(20)}$ es:

$$L_{18} = \frac{\exp(\beta_1(1) + \beta_2(3) + \beta_3(1))}{\sum_{s \in R(t_{18})} \exp(\beta_1 rx_{s(18)} + \beta_2 num_{s(18)} + \beta_3 size_{s(18)})} \tag{2.4}$$

En el numerador de (2.4) se colocan los datos de las variables del individuo al que se le calcula la verosimilitud, en el caso de los 26 sujetos se trata del individuo con $id = 13$, por lo que los valores que le corresponden son: $rx = 1$, $num = 3$, $size = 1$.

Para llevar a cabo el análisis de datos en R se puede utilizar la función **coxph** de la librería **survival**, esta función permite analizar datos del tipo recurrente. Para hacer el análisis, los datos deben estar en formato supervivencia, esto se logra como se hizo en el ejemplo de varianza robusta, con la función **Surv** la cual recibe las variables de tiempo de inicio, fin y el estatus, R por defecto sabe que 1 significa evento, por lo que en caso de que la asignación cambie se debe especificar.

En la función **coxph** se introducen los datos en formato Surv y se le asignan las covariables, recordar que para obtener varianzas robustas es necesario incluir **cluster()**:

```
datos_surv  <- Surv(inicio,stopp,event, data= bladder2)
AG <- coxph(datos_surv ~  rx +  num + size + cluster(id),
data = bladder2)
```

Tabla 2.3: Salida de R para el modelo de Andersen-Gill.

	coef	exp(coef)	se(coef)	robust se	z	$P(>\|z\|)$
rx	-0.4647	0.6283	0.1997	0.2656	-1.75	0.0801
number	0.1750	1.1912	0.0471	0.0630	2.78	0.0055
size	0.0437	2.498	0.0691	0.0776	-0.56	0.5738

Rsquare	= 0.094			
Likelihood ratio test	= 17.52 on 3 df,	p = 0.00055		
Wald test	= 11.54 on 3 df,	p=0.009122		
Score (logrank) test	= 19.52 on 3 df,	p = 0.00021,	Robust = 11.27	p = 0.01036

Como se mencionó al principio del ejemplo, la variable que es de interés es el tipo de tratamiento y las otras dos variables son reguladoras de éste. En la tabla2.3 están los resultados del ajuste del modelo Andersen-Gill, La razón de riesgo de la variable tratamiento es $\exp(coef) = \exp(-.4647) = 0.6283$, usando la prueba de Wald para probar si H_0:rx no tiene efecto. El p-value obtenido de la prueba es $0.0801 > 0.05 = \alpha$ que es mayor que el valor α de significancia, por lo que la variable es no significativa. Esto indica que el tipo de tratamiento no tiene efecto en el estudio. Por otra parte de las

Tabla 2.4: Prueba de bondad de ajuste para el modelo Andersen-Gill.

	rho	chisq	p
rx	0.02577751	0.1234454	0.7253273
number	0.06892714	0.7737622	0.3790557
size	-0.10282878	1.5227852	0.2171985
GLOBAL	NA	3.4866809	0.3224937

covariables tamaño y número únicamente el número de tumores tiene efecto en el estudio.

En la tabla 2.4 se muestran los resultados de la prueba de bondad de ajuste para las variables en el modelo de Andersen-Gill, donde se prueba si las variables que están en el modelo cumplen con el supuesto de riesgos proporcionales. Como la variable que interesa en este modelo es el tipo de tratamiento entonces se prestará especial atención en ella. A un nivel de significancia de $\alpha = 0.05$, el p-value de la variable rx es mayor que $\alpha < 0.7253$, esto indica que no se rechaza H_0, por lo que sí se cumple la hipótesis de riesgos proporcionales para esta variable. Además de hacer la prueba de bondad de ajuste, se pueden revisar los residuos Schoenfeld para tener una idea del comportamiento de las observaciones y si hubiese problemas como solucionarlos. En la figura 2.2 están los 3 gráficos de residuos Schoenfeld, un gráfico para cada variable en el modelo. Como se está estudiando el efecto del tratamiento, únicamente se observará el primer gráfico, en donde se puede ver que la línea horizontal trazada no depende de los residuos.

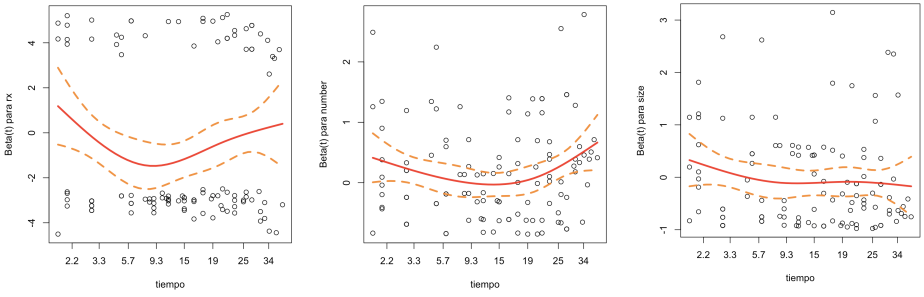

Figura 2.2: Residuos Schoenfeld para el modelo Andersen-Gill.

En la figura 2.3 se muestran los gráficos de los residuos delta-beta con los que es posible identificar si existen observaciones influyentes. En el caso de las observaciones influyentes para la variable rx, se puede ver que en efecto existe una observación influyente, que corresponde a la segunda ocurrencia del individuo 17.

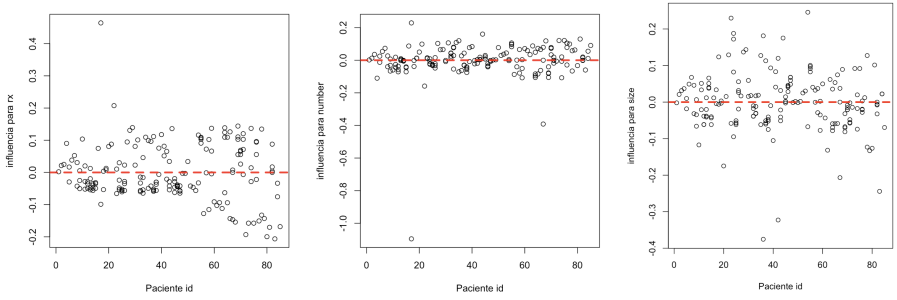

Figura 2.3: Observaciones influyentes en el modelo Andesen-Gill.

2.1.3. Modelo Prentice-William-Peterson

En el modelo Prentice-William-Peterson (PWP) el tiempo en el que ocurre el evento de estudio sí importa, ya que un individuo no estará en riesgo de que le ocurra el evento k, hasta que haya experimentado el evento $k-1$.

PWP propusieron 2 modelos de regresión para eventos recurrentes que toman en cuenta las covariables y el tiempo de falla en la función de intensidad.

En el primer modelo se toma el tiempo desde que el estudio inició hasta que ocurre el evento mientras que el segundo modelo toma en cuenta el tiempo de ínter-ocurrencia de los eventos.

Prentice, William y Peterson introdujeron el concepto de estrato, ellos definen un estrato como la cantidad de veces que un sujeto experimenta el evento de estudio. Esto es, que si el sujeto A experimenta 1 vez el evento pertenecerá al estrato 1, si el sujeto B experimenta 3 veces el evento pertenecerá al estrato 3, por lo que el estrato k contendrá a todos aquellos sujetos que hayan experimentado k veces el evento de estudio. La variable de estratificación trata al tiempo en el modelo como una variable categórica.

En consecuencia habrá tantas funciones de intensidad como estratos en el modelo, a diferencia del modelo de Andersen-Gill donde todos los eventos son iguales.

PWP presentaron 2 tipos de funciones de riesgo, la primera es una función del tiempo de inicio del estudio hasta la ocurrencia del evento t, y la otra desde el evento inmediato anterior. Entonces las funciones de riesgo de estos modelos son las siguientes:

$$\lambda_s\{t|N(t),Z(t)\} = \lambda_{0s}(t)\exp\{Z_t'(t)\beta_s\} \quad (2.5)$$

$$\lambda_s\{t|N(t),Z(t)\} = \lambda_{0s}(t-t_n)\exp\{Z_t'(t)\beta_s\} \quad (2.6)$$

para ambos casos $\lambda_{0s}(t) \geq 0$, $s = \{1, 2, ...\}$ son funciones de riesgo base arbitrarias, s es la variable de estratificación $s = \{N(t), Z(t), t\}$ que varía a través del tiempo.
$Z(t) = \{z(u) : u \leq t\}$ es el proceso de covariables hasta el tiempo t, y $Z(u) = Z_1(u), ..., Z_p(u)$ es el vector de covariables del sujeto de estudio en el tiempo $u \geq 0$. $N(t) = \{N(u) : u \leq t\}$, donde $N(u)$ es el número de eventos ocurridos hasta t. β_s es un vector columna de los coeficientes de regresión estratificados. El tiempo entre eventos se representa como $t - t_n$ donde $t < t_n$.

Se conoce como PWP1 al modelo obtenido de la función de riesgo (2.5) y PWP2 al obtenido de la función de riesgo (2.6), dado que la forma de estos dos modelos difiere en los intervalos de tiempo, al momento de introducir los datos para su estudio se debe hacer de diferente manera, para ejemplificar el diseño de datos se utilizará la información del sujeto #9 del ejemplo del modelo de AG (tabla 2.1). Como se puede observar en la figura 2.1 el sujeto #9 presentó 2 fallas y 1 censura por lo que desapareció del estudio. Los 3 eventos implican que el sujeto tenga 3 estratos, la variable que se utiliza para estratificar es *int*.

Para el caso de PWP1 se utiliza el mismo diseño de datos que para el modelo de AG a diferencia que en vez de utilizar el modelo de RP de Cox, se utiliza un modelo estratificado de Cox. En la tabla 2.5 se ilustra el diseño de datos para PWP1.

Tabla 2.5: Ejemplo sujeto 9 - PWP1.

id	**int**	event	start	stop	tx	num	size
9	1	1	0	12	0	1	1
9	2	1	12	16	0	1	1
9	3	0	16	18	0	1	1

El modelo PWP2 difiere del modelo PWP1 en el tratamiento del tiempo, por lo que a pesar de utilizar un diseño de datos con (*start*, *stop*) los tiempos que deben registrarse son distintos, esto es, el tiempo de inicio (*start*) debe ser en todos los casos 0 y el tiempo de paro (*stop*) se refiere al tamaño del intervalo de tiempo desde que ocurrió el evento anterior. Para este modelo también se debe utilizar un modelo estratificado de Cox. En la tabla 2.6 se muestra el diseño de datos para el modelo PWP2.

La principal diferencia entre PWP1 y PWP2 se encuentra en que para PWP1 el **tiempo** hasta el primer evento sí influye y afecta el conjunto en riesgo para los eventos posteriores. Mientras que en PWP2 se puede decir que "El reloj se pone en cero" cada vez que un evento ocurre, esto es que el tiempo hasta el primer evento no afecta al conjunto en riesgo.

Tabla 2.6: Ejemplo sujeto 9 - PWP2.

id	**int**	event	start	stop	tx	num	size
9	1	1	0	12	0	1	1
9	2	1	0	4	0	1	1
9	3	0	0	2	0	1	1

Ejemplo

Utilizando nuevamente los datos de R, *bladder2* se realiza el ajuste del modelo de PWP1 y PWP2. Se utiliza la misma base que para el caso del modelo Andersen Gill, pero para el modelo PWP2 se tienen que crear 2 variables extras, pues las variables de inicio y fin de los intervalos de tiempo tienen un formato distinto, tal como puede verse en la tabla 2.6.

La función de riesgo para este modelo tendría que verse de la siguiente forma:

$$\lambda_s\{t|N(t), Z(t)\} = \lambda_{0s} \exp\{\beta_s tx + \beta_{1s} num + \beta_{2s} size\} \qquad s = 1, 2, 3, 4 \tag{2.7}$$

La variable para estratificar será **int** que representa el número de fallas que ha tenido un individuo. En el caso de los modelos estraficados de Cox, debe revisarse si hay alguna interacción entre las covariables del modelo. Aún cuando no exista interacción entre las variables, en ocasiones se utiliza el modelo con interacción debido a que se puede obtener la razón de riesgos (HR) para cada estrato en el modelo. Esto se puede ver claramente en las tablas 2.7 y 2.8, en la segunda tabla no se tiene una prueba general para todos los estratos tal como en el caso de Andersen-Gill.

El modelo con interacción para el modelo PWP1 se muestra en (2.8)

$$\begin{aligned} \lambda_s\{t|N(t), Z(t)\} &= \lambda_{0s} \exp\{\beta rx + \beta_1 num + \beta_2 size \\ &+ \delta_{11}(Z_1^* \times rx) + \delta_{12}(Z_2^* \times rx) + \delta_{13}(Z_3^* \times rx) \\ &+ \delta_{21}(Z_1^* \times num) + \delta_{22}(Z_2^* \times num) + \delta_{23}(Z_3^* \times num) \\ &+ \delta_{31}(Z_1^* \times size) + \delta_{32}(Z_2^* \times size) + \delta_{33}(Z_3^* \times size)\} \\ s = 1, 2, 3, 4 \end{aligned} \tag{2.8}$$

donde Z_i^* con $i = 1, 2, 3$ es una variable Dummy para los 4 estratos, la variable escogida es *int*. La hipótesis nula en el caso del modelo de interacción para la prueba de razón de verosimilitudes se presenta en (2.9). En la tabla 2.7 se muestra la salida de R para el modelo de interacción de la ecuación (2.8),

$$\mathbf{H_0} : \delta_{11} = \delta_{12} = \delta_{13} = \delta_{21} = \delta_{22} = \delta_{23} = \delta_{31} = \delta_{32} = \delta_{33} = 0 \tag{2.9}$$

Sin embargo, no es la única forma de incluir la interacción en el modelo, pues se puede decir que para cada estrato la función de riesgo tendrá coeficientes distintos. En este caso,

al no rechazar la hipótesis nula se estaría probando que los coeficientes de las covariables son iguales para cualquier estrato.

$$\begin{aligned} \mathbf{H_0} \quad & : \beta_1 = \beta_2 = \beta_3 = \beta_4 \equiv \beta \\ & \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} \equiv \beta_1 \\ & \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} \equiv \beta_2 \end{aligned} \tag{2.10}$$

El p-value de la razón de verosimilitudes para el modelo PWP1 con interacción es menor

Tabla 2.7: Salida de R para modelo PWP1 con interacción.

	coef	exp(coef)	se(coef)	robust se	z	$P(>\|z\|)$
rx	-0.7899	0.4539	0.2732	0.2751	-2.8710	0.0041
rx1	0.8359	2.3069	0.3196	0.3865	2.1626	0.0306
rx2	1.0154	2.7604	0.4805	0.4678	2.1706	0.0300
rx3	0.8060	2.2389	0.5382	0.5104	1.5792	0.1143
size	0.0146	1.0147	0.0999	0.0915	0.1591	0.8736
size1	-0.0795	0.9236	0.1367	0.1459	-0.5445	0.5861
size2	0.2470	1.2802	0.1988	0.1486	1.6626	0.0964
size3	0.0256	1.0259	0.2227	0.2146	0.1193	0.9050
number	0.2272	1.2551	0.0777	0.0751	3.0255	0.0025
number1	-0.2043	0.8152	0.1141	0.1525	-1.3400	0.1802
number2	-0.0706	0.9318	0.1812	0.1653	-0.4272	0.6693
number3	0.0344	1.0350	0.1854	0.1571	0.2187	0.8269

Rsquare	= 0.273			
Likelihood ratio test	= 56.68 on 12 df,	p=9.039e-08		
Wald test	= 80.34 on 12 df,	p=3.559e-12		
Score (logrank) test	= 69.51 on 12 df,	p=3.963e-10	Robust = 26.06	p=0.01052

que el nivel de significancia $9.039e-08 < \alpha = 0.05$ por lo que se rechaza la hipótesis nula de (2.9). En el caso del modelo sin interacción el p-value es mayor que el nivel de significancia $0.06006 > \alpha = 0.05$, lo que indica que no se rechaza la hipótesis nula H_0: $\beta_1 = \beta_2 = \beta_3 = 0$.

En la tabla 2.9 se muestra el ajuste del modelo PWP2, a partir de estos resultados se realiza la prueba de riesgos proporcionales. En este modelo la única variable significativa es *number*.

Note que bajo el modelo PWP1 las variables *rx* y *number* son significativas, mientras que en el modelo PWP2 solo *number* es significativa.

Tabla 2.8: Salida de R para modelo PWP1.

	coef	exp(coef)	se(coef)	robust se	z	$P(>\|z\|)$
rx	-0.43228	0.64903	0.22082	0.21655	-1.996	0.0459
number	0.11557	1.12252	0.05373	0.05340	2.164	0.0304
size	-0.01566	0.98446	0.07297	0.06219	-0.252	0.8011

Rsquare	= 0.041			
Likelihood ratio test	= 7.4 on 3 df,	p=0.06006		
Wald test	= 7.49 on 3 df,	p=0.0578		
Score (logrank) test	= 7.71 on 3 df,	p=0.0523,	Robust = 8.83	p=0.0316

Tabla 2.9: Salida de R para modelo PWP2.

	coef	exp(coef)	se(coef)	robust se	z	$P(>\|z\|)$
rx	-0.27900	0.75653	0.207347	0.21562	-1.29394	0.19568
number	0.15804	1.17121	0.05194	0.05093	3.10260	0.00191
size	0.00741	1.00744	0.07002	0.06433	0.11526	0.90823

Rsquare	= 0.051			
Likelihood ratio test	= 9.33 on 3 df,	p=0.02517		
Wald test	= 11.84 on 3 df,	p=0.00795		
Score (logrank) test	= 10.27 on 3 df,	p=0.01640,	Robust = 9.92	p=0.0192276179

En la tabla 2.10 se muestra la prueba de bondad de ajuste para los modelos de PWP, en ambos casos la variable de tratamiento *rx* cumple con el supuesto de riesgos proporcionales. La figura 2.4 contiene los residuos Schoenfeld para el modelo PWP1, en dichos gráficos puede verse que la linea trazada horizontalmente no depende de las observaciones. Por otra parte en la figura 2.5 se pueden ver las observaciones influyentes representadas con los residuos delta-beta, en este modelo, claramente la observación del individuo 20 es influyente para la variable *number*.

En la figura 2.6 se muestran los gráficos de los residuos Schoenfeld para el modelo PWP2, puede verse claramente que la variable *number* tiene una línea recta horizontal lo que confirma que se cumple el supuesto de riesgos proporcionales. En la figura 2.7 se muestran los residuos delta-beta con los que se comprueba si existen variables influyentes. El individuo 19 es influyente para las variables *number* y *rx*

2.1.4. Modelo Wei-Lin-Weissfeld

El modelo Wei-Lin-Weissfeld es una extensión del modelo de Cox de riesgos proporcionales, y fue desarrollado para tratar los casos de tiempo de falla multivariados.

Tabla 2.10: Prueba de bondad de ajuste para modelos PWP1 y PWP2.

Modelo PWP1

	rho	chisq	p
rx	0.0494	0.242	0.623
number	-0.0623	0.400	0.527
size	-0.0630	0.330	0.566
GLOBAL	NA	0.714	0.870

Modelo PWP2

	rho	chisq	p
rx	-0.14834	2.561	0.109
number	0.00955	0.009	0.920
size	-0.06220	0.414	0.520
GLOBAL	NA	2.680	0.444

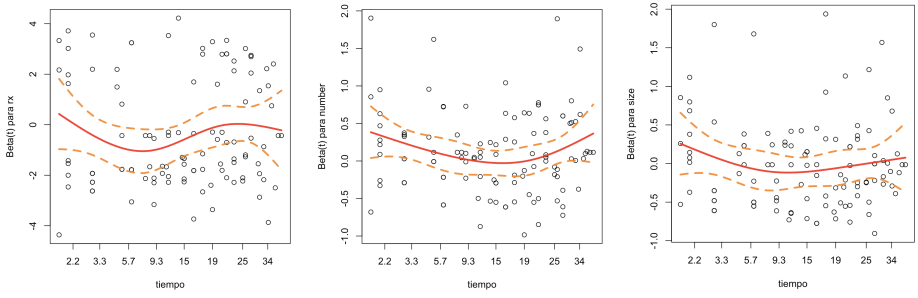

Figura 2.4: Residuos Schoenfeld para el modelo PWP1.

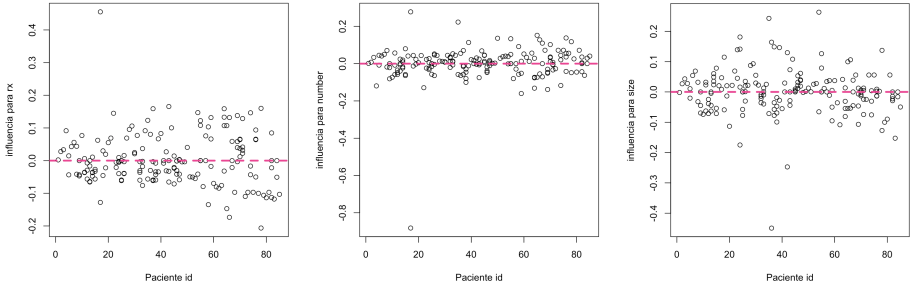

Figura 2.5: Identificación de observaciones influyentes en el modelo PWP1.

En este modelo todos los sujetos estarán en todos los estratos aún cuando no hayan presentado todos el máximo de eventos, es por esto que habrá tantos estratos como eventos ocurridos. Supóngase que en un estudio el máximo número de eventos que un mismo individuo presentó fue K, en consecuencia todos los individuos en el estudio, aun los que solo tuvieron 1 evento (falla o censura) estarán presentes en los K estratos. En este último caso todos los estratos del individuo que presentó 1 evento serán idénticos al anterior.

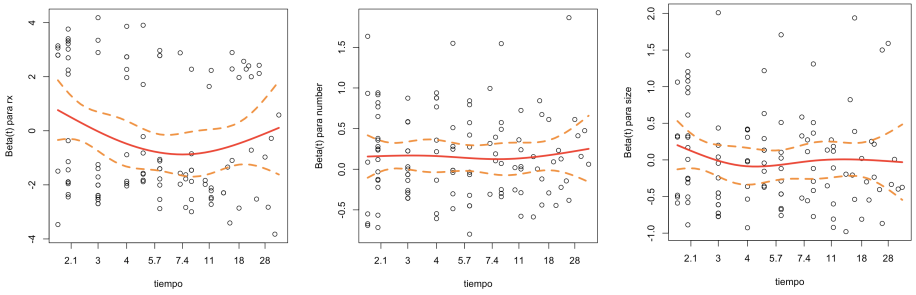

Figura 2.6: Residuos Schoenfeld para el modelo PWP2.

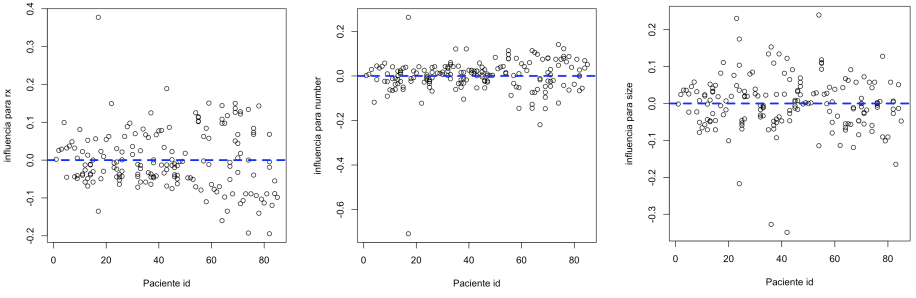

Figura 2.7: Identificación de observaciones influyentes en el modelo PWP2.

La función de riesgo del i-ésimo individuo, al momento de la j-ésima ocurrencia es:

$$\lambda_{ij}(t) = Y_{ij}(t)\lambda_{0j}(t)\exp\{Z'_{ij}\beta_j\} \tag{2.11}$$

T_{ij} es el tiempo de la j-ésima ocurrencia del evento sobre el sujeto i, con $j = 1, ..., k$ e $i = 1, ..., n$. El máximo número de eventos ocurridos durante el estudio se representa con k.

$Z_{ij} = (Z_{ij,1}(t), Z_{ij,2}(t), ..., Z_{ij,p}(t))$ es el vector de covariables para el sujeto i al tiempo t respecto al evento j.

Y_{ij} es una variable indicadora igual a 1 hasta la ocurrencia del evento j, excepto cuando el sujeto esté censurado.

La función de verosimilitud parcial para el s-ésimo estrato $s = 1, .., k$ está dada por:

$$L_s(\beta_s) = \prod_{i=1}^{n}\left[\frac{\exp(\beta'_s Z_{si})}{\sum_{l\in R_s(t_{si})}\exp(\beta'_s Z_{sl})}\right]$$

A diferencia del modelo de Andersen-Gill, el diseño de los datos en el modelo WLW no contiene las columnas de *start* y *stop*, esto se ejemplifica en la tabla 2.11 utilizando

los datos del ejemplo del modelo de AG del individuo 9.

Tabla 2.11: Ejemplo sujeto 9

id	int	event	stime	tx	num	size
9	1	1	12	0	1	1
9	2	1	16	0	1	1
9	3	0	18	0	1	1
9	4	0	18	0	1	1

Por otra parte se puede observar que a pesar de que el sujeto 9 únicamente presentó 2 fallas tiene 4 estratos, esto ocurre debido a que el mayor número de intervalos presentados es 4 (sujeto #14 y #25).

Ejemplo

Para el caso de un modelo WLW, como se mencionó anteriormente, únicamente se toma en cuenta el tiempo *stop* y el inicio siempre será 0, por lo que se debe hacer una modificación al diseño de datos. Para realizar este modelo se utilizará la base *bladder* que contiene ya los datos en el formato que se necesita. En la tabla 2.12 se muestra la salida de R.
Se puede ver que en el caso del modelo WLW la variable del tratamiento rx es ligeramente no significativa.

Tabla 2.12: Salida de R para modelo WLW.

	coef	exp(coef)	se(coef)	robust se	z	$P(>\|z\|)$
rx	-0.5847	0.5572	0.2010	0.3079	-1.899	0.0575
number	0.2102	1.2340	0.0467	0.0666	3.155	0.0016
size	-0.05161	0.9496	0.0697	0.0945	-0.5457	0.5852

Rsquare	= 0.072			
Likelihood ratio test	= 25.26 on 3 df,	p=1.3618e-05		
Wald test	= 15.54 on 3 df,	p=0.00141		
Score (logrank) test	= 28.6 on 3 df,	p=2.718e-06,	Robust = 11.63	p=0.00876

La prueba de bondad de ajuste del modelo WLW presentada en la tabla 2.13 para la variable explicativa rx muestra un p-value de $0.6482 > \alpha = 0.05$ mayor que el nivel de significancia, esto implica que H_0 no se rechaza, por lo que sí se cumple el supuesto de riesgos proporcionales sobre esta variable. Esto puede además reafirmarse si se ve la figura 2.8. En el primer gráfico se tienen los residuos Schoenfeld de la variable tratamiento

Tabla 2.13: Prueba de bondad de ajuste para WLW.

	rho	chisq	p
rx	0.0281	0.208	0.6482
number	0.0784	1.194	0.2744
size	-0.1227	3.172	0.0749
GLOBAL	NA	5.845	0.1194

rx, en este gráfico se puede observar que la línea trazada si bien no es constante, no se mueve conforme los datos observados. Con lo que se confirma que se cumple el supuesto de riesgos proporcionales. En la figura 2.9 se muestran las observaciones influyentes para el modelo WLW, en este caso tanto para la variable *number* u para la variable rx el sujeto 19 resulta influyente

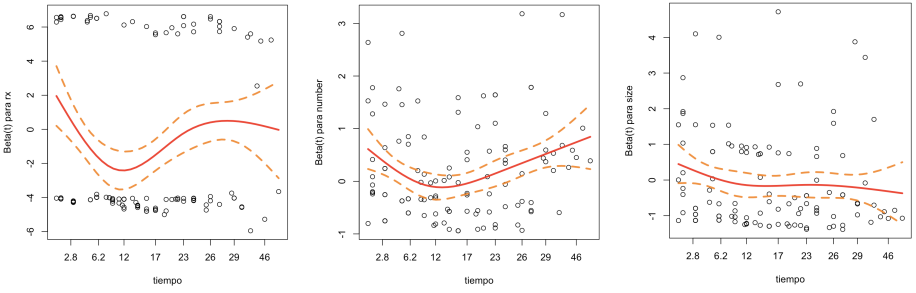

Figura 2.8: Residuos Schoenfeld para el modelo WLW.

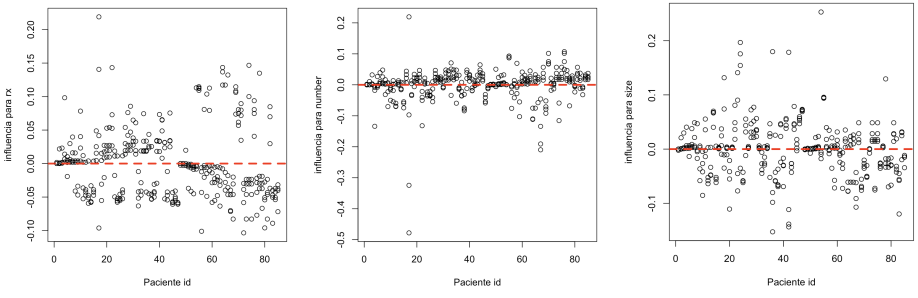

Figura 2.9: Observaciones influyentes en el modelo WLW.

Finalmente en la tabla 2.14 se presenta la comparación de los parámetros estimados de los 4 modelos vistos anteriormente, para así poder concluir si la variable de tratamiento influye en la aparición de nuevos tumores controlada por el número de tumores iniciales

y tamaño.
Únicamente el modelo de AG es un modelo de Cox de riesgos proporcionales estándar, los demás son estratificados. El p-value de la prueba de Wald para todos los modelos indica que solamente el modelo PWP1 es significativo. En la figura 2.10 y en 2.11 se muestran los gráficos de la función de riesgo para cada uno de los modelos. Tanto en los modelos de PWP como en el de WLW se tienen 4 funciones de riesgo pues hay 4 estratos en cada modelo.

Tabla 2.14: Comparación de parámetros estimados variable rx.

Modelo	$\hat{\beta}$	$\hat{HR} = \exp(\hat{\beta})$	p-value
AG	-0.46469	0.62833	0.08015
PWP1	-0.43228	0.64903	0.0459
PWP2	-0.279005	0.756536	0.19569
WLW	-0.58479	0.55722	0.0576

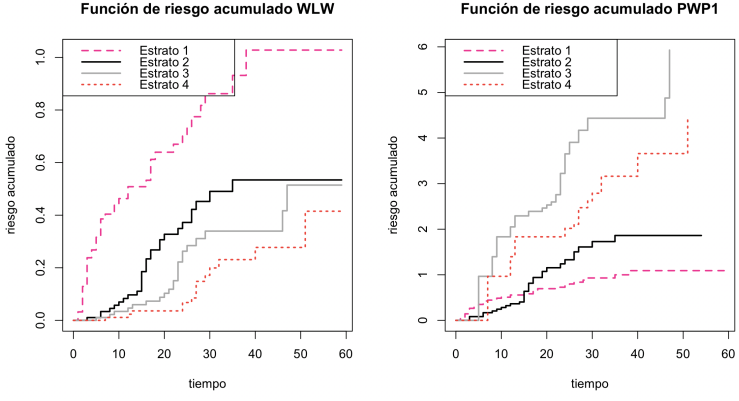

Figura 2.10: Funciones de riesgo acumulado

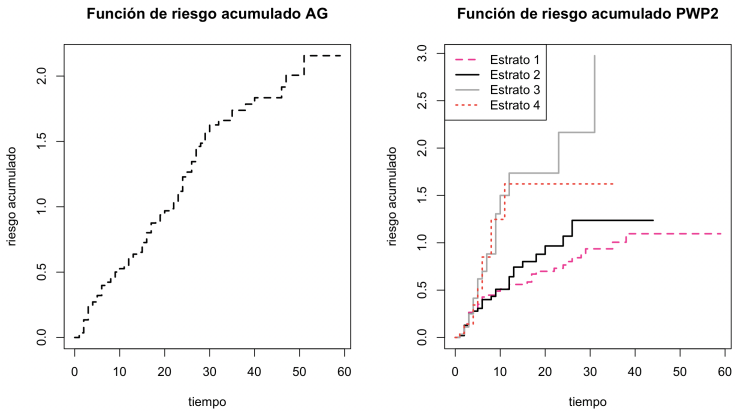

Figura 2.11: Funciones de riesgo acumulado

2.2. Modelos No Paramétricos

En los últimos años, se han desarrollado modelos no paramétricos para modelar los eventos recurrentes. Dentro de estos modelos se encuentran el desarrollado por Wang-Chang (1999) y el de Peña, Strawderman y Hollander (2001).

2.2.1. Modelo Wang- Chang

El modelo de Wang-Chang (WC) permite estimar curvas de supervivencia con eventos recurrentes tanto para los casos donde los tiempos entre ocurrencias son independientes, como para los casos donde los datos están correlacionados.

El estimador que proponen WC se basa en dos procesos de conteo $d^*(t)$ y $R^*(t)$. Donde $d^*(t)$ es el número de individuos con tiempos de ínter-ocurrencia iguales a t, cuando al menos un evento ocurre y $R^*(t)$ es el promedio de individuos que están en riesgo al tiempo t. El estimador de la función de supervivencia se muestra en la ecuación (2.12),

$$\hat{S}(t) = \prod_{i=1}^{n} \prod_{j:T_{ij} \leq t} \left[1 - \frac{d^*(T_{ij})}{R^*(T_{ij})} \right] \tag{2.12}$$

donde T_{ij} es el j-ésimo tiempo de ínter-ocurrencia del evento del i-ésimo individuo y n es el número total de tiempos de ínter-ocurrencia.

En la figura 2.12 se muestra gráficamente la recurrencia para el sujeto i-ésimo. Donde S_{ij} representa el j-ésimo tiempo calendario de la j-ésima ocurrencia del individuo i-ésimo. El $S_{i0} = 0$ ya que aún no ha sucedido ningún evento cuando inicia el estudio y $S_{ij} = T_{i1} + T_{i2} + ... + T_{iki}$. K_i es el número máximo de eventos que presenta el i-ésimo

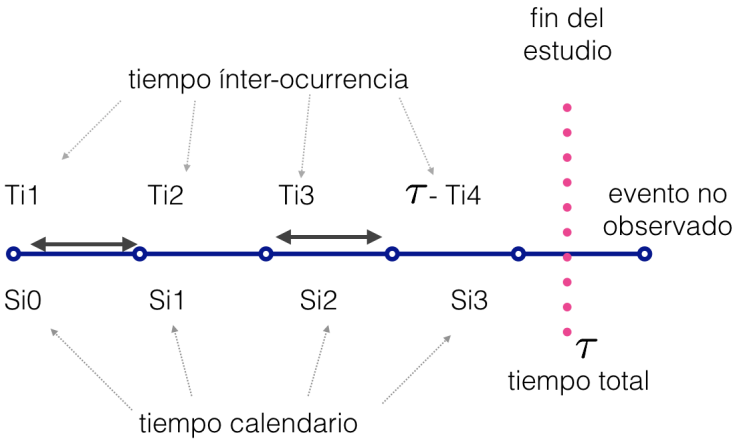

Figura 2.12: Recurrencia para el i-ésimo individuo

sujeto, $K_i = \text{máx}\{j : S_{ij} \leq \tau_i\}$, donde τ_i es el tiempo total durante el que se observó al individuo.

El proceso de conteo $d^*(t)$ se calcula de la siguiente forma:

$$d^*(t) = \sum_{i=1}^{n} \left\{ \frac{I\{K_i > 0\}}{K_i^*} \sum_{i=1}^{K_i} I\{T_{ij} = t\} \right\}$$

donde

$$K_i^* = \begin{cases} 1 & \text{si } K_i = 0 \\ K_i & \text{si } K_i > 0 \end{cases}$$

por otra parte,

$$R^*(t) = \sum_{i=1}^{n} \frac{1}{K_i^*} \left[\sum_{j=1}^{K_i} I\{T_{ij} \geq t\} + I\{\tau_i - S_{iK_i} \geq t\} I\{K_i = 0\} \right]$$

2.2.2. Modelo Peña, Strawderman y Hollander

Peña, Strawderman y Hollander (PSH) propusieron un modelo para trabajar los eventos de tipo recurrente tratando de resolver 3 problemas iniciales.

- Conocer la distribución del tiempo al primer evento: esto es debido a que normalmente el tiempo entre eventos posteriores al primero es distinto.

- Resolver la correlación entre tiempos de ínter-ocurrencia.
- Resolver el efecto de las covariables en los tiempos de ínter-ocurrencia.

En el modelo de PSH, se desarrolla un estimador no paramétrico donde todos los individuos están idénticamente distribuidos, este estimador se considera como una generalización del estimador de Kaplan-Meier.

Para este modelo se utilizan 2 procesos de conteo doblemente indexados Y y N, esto es debido a que se miden con 2 escalas de tiempo distintas, la primera es la escala S de tiempo calendario y mide el tiempo que transcurre entre el inicio del estudio hasta un tiempo t. La otra escala es T la cual mide el tamaño de los tiempos ínter-ocurrencias.

El primer proceso de conteo está definido como $N(s,t)$, este proceso mide los eventos con tiempos de ínter-ocurrencia menores o iguales a t, por lo que el proceso $N_i(s,t)$ mide el tiempo del i-ésimo individuo bajo estudio.
El proceso $Y(s,t)$ mide el número de eventos ínter-ocurrencia mayores que t, ambos procesos se miden en el intervalo $[0,s]$. Si se tienen n sujetos en el estudio, todos independientes entre si, T_ij será el tiempo que ocurre entre el evento $(j-1)$ y el j para todo $i = 1,2,...,n$ y $j = 1,2,...,K_i$, donde $K_i = \text{máx}\{j : j \in \{0,1,2,...\} : S_{iK_i}\}$, esto quiere decir, que K_i es el número máximo de veces que el individuo i-ésimo presenta el evento de interés en el intervalo $[0,\tau_i]$, donde τ_i es el tiempo total que se observó a dicho individuo

$$N(s,t) = \sum_{i=1}^{n} N_i(s,t)$$
$$Y(s,t) = \sum_{i=1}^{n} Y_i(s,t)$$

donde,

$$Y_i(s,t) = \sum_{j=1}^{K_i(s-)} I\{T_{ij} \geq t\} + I\{\text{mín}(s,\tau_i) - S_{iK_i(s-)} \geq t\} \qquad \forall i = 1,2,...,n$$
$$N_i(s,t) = \sum_{j=1}^{K_i(s-)} I\{T_{ij} \leq t\} \qquad \forall i = 1,2,...,n$$

Como el estimador de la función de supervivencia desarrollado por PSH es una generalización del estimador Kaplan-Meier se necesita definir dos conceptos, que serán utilizados en dicha generalización.

$$N(s,\Delta w) = N(s,t+\Delta w) - N(s,w) \tag{2.13}$$

La ecuación (2.13) representa el número de eventos que ocurrieron en el intervalo $[0,s]$ con tiempos de ínter-ocurrencia **exactamente** iguales a w. Este término representa el

d_j del estimador KM.

El siguiente concepto necesario para el estimador es $Y(s, w)$, el proceso cuenta el número de eventos que han sucedido durante $[0, s]$ mayores o iguales al tiempo w. Lo que indica el proceso es: cuántos sujetos sobreviven al tiempo w.
Con esto se define el estimador de PSH como:

$$\hat{S}(t) = \prod_{w \leq t} \left[1 - \frac{N(s, \Delta w)}{Y(s, w)} \right]$$

2.3. Aproximación Paramétrica vía Modelo Frailty Compartido

Los modelos *frailty* compartida, también conocidos como modelos de efectos aleatorios, pueden aplicarse a los eventos recurrentes, por el hecho de que en ocasiones los múltiples eventos sobre un mismo individuo no son independientes.

En un modelo *frailty* se introduce una covariable aleatoria en el modelo con la cual se busca agregar la dependencia entre los eventos recurrentes. Este efecto aleatorio, muestra el exceso de riesgo para los individuos distintos.

Un ejemplo de un efecto aleatorio se puede ver al analizar los datos de un estudio de infecciones recurrentes donde todos los individuos tienen distintos riesgos que no son medidos mediante las covariables del modelo.

Para ejemplificar cómo funciona el efecto aleatorio compartido sobre eventos recurrentes se utilizará la misma base de cáncer de vejiga que se usó en los modelos semiparamétricos. Para esto se supondrá que la función de riesgo no condicional sigue una distribución Weibull y la fragilidad se distribuye Gamma,

$$h_i(t|\alpha, X_i) = \alpha_i h(t|X_i) \tag{2.14}$$

donde $\alpha \sim \text{gamma}(media = 1, varianza = \theta)$ y además $h(t|X_i) = \lambda_i p t^{p-1} \exp\{\beta X_i\}$ tal como en la ecuación 1.42.

Existe una paquetería en R llamada *frailtypack* con la cual se puede llevar a cabo el análisis de datos recurrentes mediante fragilidad compartida. En [Król et al., 2017],[Rondeau and Gonza puede consultarse a detalle sobre los modelos de fragilidad y la aplicación en R, son de mucha utilidad si se quiere profundizar en el tema.

En R el comando para realizar un modelo de efectos aleatorios es el siguiente:

```
recu_shafrai <-frailtyPenal(Surv(start,stop,event)~
                      cluster(id)+rx+size+number,
                      data=bladder2,recurrentAG=TRUE,
                      hazard=\Weibull")
```

Tabla 2.15: Estimación de los parámetros del modelo de fragilidad compartida gamma mediante la aproximación paramétrica de la función de riesgo.

	Coef	exp(coef)	SE coef (H)	z	p
rx	-0.6125497	0.541967	0.3296975	-1.857914	0.0631810
size	-0.0147194	0.985388	0.1117133	-0.131761	0.8951700
number	0.2480898	1.281575	0.0938931	2.642258	0.0082355

Parámetro fragilidad Theta	1.05032	(SE (H): 0.350028)	p = 0.0013469
Parámetro escala función de riesgo Weibull	14.64		
Parámetro forma función de riesgo Weibull	1.08		

En la tabla 2.15 se muestra la salida de R del ajuste del modelo con fragilidad. La razón de riesgo estimada de la variable rx obtenida es $\exp(coef) = \exp(-0.6125497) = 0.541967$, este valor se obtiene de comparar 2 sujetos con la misma fragilidad. Este valor es parecido al que se obtuvo en el modelo WLW. Por otra parte el p-value de θ estimado resulta ser significativo, lo que quiere decir que el componente de fragilidad sí influye en el modelo y los sujetos tienen cierta correlación en sus observaciones.

La diferencia entre la varianza robusta que se utilizó en los modelos semiparamétricos y el modelo *frailty* es que esta última influye en la estimación de los parámetros.

Conclusiones

Ahora que se conocen los tres modelos semiparamétricos de eventos recurrentes las preguntas que surgen naturalmente son ¿qué modelo se debe usar?, ¿cuál es el mejor? y la respuesta es: depende del objetivo del estudio.

El modelo de Andersen-Gill (AG) se debe utilizar cuando todos los datos son tratados de la misma manera, sin importar el tiempo de ocurrencia de los eventos, este modelo ofrece la ventaja de ser más sencillo de realizar. La desventaja de utilizar un modelo AG es que, en caso de que los eventos sean dependientes, los errores estándar estarán subestimados lo que ocasionaría tener un mayor número de errores del tipo I. Una solución sería modelar con covariables dependientes del tiempo.

Si los eventos son de distinto tipo, se recomienda usar un modelo Wei-Lin-Weissfeld (WLW), en el caso del ejemplo de cáncer de vejiga, todos los eventos son iguales, por lo que se recomendaría utilizar AG.

Ahora, si lo importante en el estudio es estimar el efecto de la variable según el orden de los eventos se recomienda utilizar un modelo Pretince-William-Peterson (PWP), estos modelos indican que un individuo solo podrá estar en riesgo de presenciar otra falla si ya le ocurrió la anterior. Un modelo marginal (WLW) debería utilizarse cuando el interés recae en estimar el número esperado de eventos o la tasa de recurrencia condicionada a las covariables. Si lo que importa al modelar los datos es el tiempo ínter-ocurrencia se escoge un modelo PWP2, en cambio si esto no importa y solo se quiere tomar en cuenta el orden se utiliza un modelo PWP1.

Los modelos de AG y PWP desarrollados a través del capítulo 2 asumen que los eventos pasados dependen solo del pasado inmediato, mientras que los modelos *frailty* suponen dependencia en los eventos a través de los efectos aleatorios compartidos. Por otra parte, los modelos marginales WLW permiten dependencia entre los eventos, al caracterizar las tasas mediante el proceso de conteo.

Un modelo *frailty* debe utilizarse cuando quiera tomarse en cuenta que existen factores aleatorios, no observados, que están afectando el modelo. El modelo *frailty* utilizado en este trabajo es una extensión del modelo de Cox de riesgos proporcionales multiplicativo, sin embargo no es el único modelo *frailty* existente.

Los resultados de la razón de riesgos (HR) para el modelo Andersen-Gill muestran

que después del primer evento se reduce un 38 % el riesgo de recurrencia. Mientras que el modelo *frailty* muestra que cada tumor nuevo incrementa el riesgo de recurrencia en un 28 % ya que su razón de riesgo es HR=1.28, sin embargo comparar todos los modelos no es recomendable pues cada uno se ocupa con distintos objetivos.

Bibliografía

[Amorim and Cai, 2015] Amorim, L. D. and Cai, J. (2015). Modelado de eventos recurrentes: un tutorial para el análisis en epidemiología. *International Journal of Epidemiology*, 44(1):324–333.

[Collet, 2015] Collet, D. (2015). *Modelling Survival Data in Medical Research*. CRC Press, 3rd edition.

[Cook and Lawless, 2007] Cook, R. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer, 2nd edition.

[Cárdenas Leuro, 2013] Cárdenas Leuro, M. (2013). Un modelo de sobrevida multivariado para eventos recurrentes por sujeto con evento terminal: deserción de clientes en la industria de las telecomunicaciones. Master's thesis, Universidad Nacional de Colombia , Bogotá,Colombia.

[Kleinbaum and Klein, 2012] Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis A Self-Learning Text*. Springer, 3rd edition.

[Król et al., 2017] Król, A., Mauguen, A., Mazroui, Y., Laurent, A., Michiels, S., and Rondeau, V. (2017). Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software*, 81(3):1–52.

[Martinez, 2009] Martinez, C. (2009). *Generalización de algunas pruebas clásicas de comparación de curvas de supervivencia al caso de eventos de naturaleza recurrente*. PhD thesis, Universidad Central de Venezuela.

[Moeschberger and Klein, 2003] Moeschberger, M. L. and Klein, J. P. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. Springer, 2nd edition.

[Moore, 2016] Moore, D. F. (2016). *Applied Survival Analysis Using R*. Springer, 1st edition.

[R Core Team, 2017] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rebaza Fernández, 2017] Rebaza Fernández, D. d. R. (2017). Modelos semiparamétricos de eventos recurrentes: caso aplicación a pacientes con cáncer de mama. Master's thesis, Universidad Nacional Agraria La Molina.

[Rondeau and Gonzalez, 2005] Rondeau, V. and Gonzalez, J. R. (2005). Frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer Methods and Programs in Biomedicine*, 80(2):154–164.

[Therneau and Grambsch, 2000] Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, 1st edition.

[Xian, 2012] Xian, L. (2012). *Survival Analysis: Models and Applications*. Wiley, 1st edition.