UNIVERSIDAD AUTÓNOMA METROPOLITANA
DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

Coeficiente de Gini para datos censurados de duración del desempleo

Tesis

para obtener el grado de:

Maestro en Ciencias (Matemáticas Aplicadas e Industriales)

presenta:

Oswaldo Guevara Munive

Asesora: Dra. Blanca Rosa Pérez Salvador

Jurado calificador:

Presidente:	**Dr. José Raúl Montes de Oca Machorro**	UAM-I
Secretario:	**Dra. Blanca Rosa Pérez Salvador**	UAM-I
Vocal:	**Dr. Hugo Adán Cruz Suárez**	BUAP

Lugar: Edificio de Posgrado EP-001, Universidad Autónoma Metropolitana-Iztapalapa
Fecha: 18 de Diciembre de 2019 a las 13:00 hrs.

DEDICATORIA

A mi madre, porque mis logros son también resultado de su esfuerzo, de su esfuerzo por mi bienestar de toda la vida.

Agradecimientos

A mi madre más que a nada ni nadie por su apoyo incondicional y esfuerzo incansable.

A mi familia que me ha brindado siempre amor y apoyo.

A la UAM-Iztapalapa por darme cabida y dejarme ser parte de su comunidad.

A la Doctora Blanca Rosa Pérez Salvador, mi asesora, por su paciencia y trabajo para el buen término del trabajo.

A la Maestría de Ciencias Matemáticas Aplicadas e Industriales por darme la confianza de realizar estos estudios.

A mis profesores de la maestría por su trabajo profesional.

A mis sinodales, el Doctor Hugo Adán Cruz Suárez y el Doctor Raúl Montes de Oca, por su apoyo y comprensión.

Índice general

Resumen

El trabajo busca llevar a la práctica para un caso mexicano los estimadores del coeficiente de Gini propuestos por Lv, G. Zhang y Ren 2017. Para ello analizamos primero dichos estimadores, teóricamente a través del Análisis de Supervivencia pero sobre todo por medio de la teoría de Procesos de Conteo Martingala, y también por medio de simulación computacional (con el uso del Software R) probamos los estimadores bajo diferentes condiciones y diferentes distribuciones de probabilidad. Por último aplicamos esta propuesta al caso de los egresados de posgrado en México para abordar el tema de la desigualdad en el tiempo para encontrar trabajo.

Introducción

Poder medir la desigualdad en la Ciencia Económica es un problema práctico de primera importancia. Puede ser considerado como un indicador del éxito de una economía si en ella también se ha combatido exitosamente la pobreza. La medida más conocida, más desarrollada y utilizada es el coeficiente de Gini.

Un caso particular de entre sus muchos enfoques y objetivos es la medición de desigualdad para una variable con algún tipo de censura, es decir que las mediciones no son completas por algún motivo o que hubo algún tipo de restricción para realizar las observaciones (mediciones). Uno podría desechar las observaciones censuradas y hacer los cálculos sólo con aquellas observaciones no censuradas pero la idea detrás de la incorporación de los elementos censurados es que ellos guardan información relevante al fenómeno que se estudia, sin la cual los análisis serían sesgados.

Frecuente puede ser la presencia de observaciones censuradas si nuestra variable de interés es el tiempo hasta la ocurrencia de un evento. Para los economistas, la duración del desempleo es un aspecto que puede reflejar el éxito de la economía nacional o, por el contrario, un problema en ella. A nivel agregado se suele reportar la duración media del desempleo pero este indicador no dice nada sobre las diferencias entre los desempleados ni podría por sí mismo servir para concluir la existencia de un problema, ya que el promedio de la duración podría verse "movido" hacia una cantidad "pequeña" o "grande" por unas pocas observaciones atípicas. Tampoco podría señalar que la diferencia entre las mediciones de los sujetos son algo suficientemente importante que valdría la pena analizar las causas de esas diferencias.

El trabajo se basa en el reciente artículo de Lv, G. Zhang y Ren 2017, cuya propuesta y contribución al desarrollo los estimadores del coeficiente de Gini con censura es el uso del método de Ponderación de Probabilidad Inversa que, como su nombre lo indica, consiste en usar la inversa de la probabilidad de las observaciones como ponderador para corregir, en la medida de lo posible, el sesgo que provocan los datos censurados.

Con esta misma idea los autores proponen propone además un estimador para censura dependiente de covariables. El tiempo de ocurrencia del evento de interés (conseguir un empleo) y el tiempo de censura son condicionalmente independientes si son explicados por un vector de covariables y sus distribuciones de probabilidad son independientes dado el vector de covariables.

Nuestro objetivo aplicar estas ideas a la situación mexicana ya que esta evaluación nos parece de valor para detectar un problema importante además de que sería una de las primeras evaluaciones del desempleo mexicano con este enfoque, bajando del nivel agregado nacional al análisis más directo y concreto de los ciudadanos.

Para ello, habrá que hacer un análisis de tales herramientas para conocer sus propiedades, sus posibles debilidades o fortalezas, y los límites de su aplicación que nos indican cuando pueden o no utilizarse.

Capítulo 1

Índice de Gini y Análisis de Supervivencia

1.1. Índice de Gini

El índice o coeficiente de Gini apareció por primera vez en 1912, en el libro "Variabilidad y Mutabilidad" del italiano Corrado Gin. La formulación del índice de Gini pertenece a la primera parte del libro, la que trata la Variabilidad, donde él lo definió como "la diferencia media de todas las cantidades", a diferencia de otras medidas similares del momento (Ceriani y Verme 2012).

De acuerdo con este mismo trabajo de Ceriani y Verme 2012, Gini planteaba que la variabilidad se analizaba según el tipo de objeto que se iba a medir, para algunos objetos, en los que la medida era difícil y variaba de una medición a otra, el análisis tendría que responder: ¿qué tanto varían las mediciones del valor real? Mientras que en el análisis de objetos que pudieran medirse con precisión, el objetivo sería determinar: ¿qué tanto varían los objetos unos de otros? Así, la primera formulación de su índice fue llamada "diferencia media entre n cantidades" y se expresaba como

$$\Delta = \frac{2}{n\,(n-1)} \sum_{i=1}^{(n+1)/2} (n+1-2i)\,(a_{n-i+1} - a_i)$$

donde a_i son cantidades en orden ascendente, con $i = 1, 2 \ldots n$.

El índice Gini es hoy en día la medida más popular de desigualdad, utilizada sobre todo en Economía y Ciencias Sociales. En su formulación más conocida, este índice se relaciona con la curva de Lorenz.

1.1.1. La curva de Lorenz

La curva de Lorenz fue formulada por el economista estadounidense Max Otto Lorenz en 1905 para describir la desigualdad del ingreso.[1] La curva de Lorenz representa gráficamente la repartición de una cantidad entre los elementos de una población y representa esta relación proporcionalmente, de modo que cada punto de la curva se interpreta como el porcentaje o la proporción acumulada de la población que le corresponde una proporción acumulada de la variable en cuestión.

Una de las ecuaciones que define la curva de Lorenz es

$$L(t) = \frac{\int_{-\infty}^{t} uf(u)\,du}{\int_{-\infty}^{\infty} uf(u)\,du} = \frac{\int_{-\infty}^{t} uf(u)\,du}{\mu} \tag{1.1}$$

donde t es un valor dado de la variable aleatoria T, con función de distribución acumulada $F(t)$.

Una curva de Lorenz representada en una recta con pendiente positiva igual a 1 indicaría que la distribución es igual para toda la población porque el porcentaje de la población acumulado aumenta a la misma medida que la proporción acumulada de la variable repartida o distribuida.

El coeficiente de Gini se define como el cociente que resulta de dividir dos áreas: el área entre la curva ideal (recta con pendiente igual a 1) y la curva empírica (curva construida con las mediciones), y el área total debajo de la curva ideal (1/2).

[1] Nota revisada en:
Colaboradores de Wikipedia. (2018, 22 de septiembre). Max O. Lorenz. En Wikipedia, The Free Encyclopedia. Consultado a las 16:05, 30 de noviembre de 2018, de https://en.wikipedia.org/w/index.php?title=Max_O._Lorenzoldid=860706437

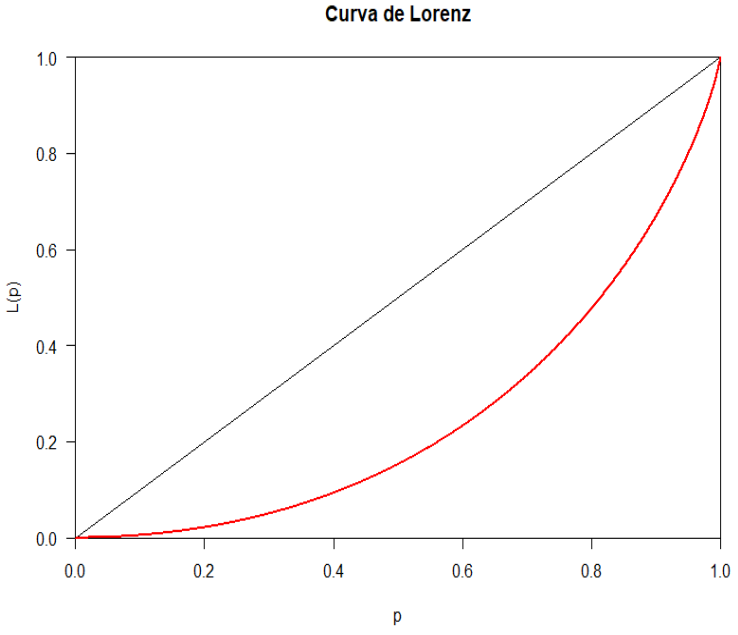

Figura 1.1: Ejemplo de la curva de Lorenz en color rojo.

Como en Gastwirth 1972, dado un conjunto de n observaciones ordenadas, $t_1 \leq t_2 \leq \ldots \leq t_n$, la curva empírica de Lorenz se forma por los puntos $\frac{i}{n}$, donde $i = 0, \ldots, n$, $L(0) = 0$ y $L\left(\frac{i}{n}\right) = \frac{s_i}{s_n}$, con $s_i = t_1 + \ldots + t_i$.

La curva empírica de Lorenz, $L(p)$, está definida para toda p, por interpolación lineal, en el intervalo $(0, 1)$ y representa la proporción de la cantidad total medida de la variable de interés perteneciente a la proporción $p - ésima$ más pequeña de la población. Suponemos que las observaciones t_i pertenecen a una distribución de probabilidad así que $0 < F(t) \leq 1$ y la media μ existe.

Definimos el coeficiente de Gini como

$$G = 1 - 2\int_0^1 L(p)dp \tag{1.2}$$

Proposición 1. *El coeficiente de Gini dado en (1.2) es equivalente a*

$$G = \frac{\int_0^\infty 2F(t)tf(t)dt}{\mu} - 1$$

Demostración:

Partiendo de la ecuación (1.1) y realizando la sustitución $F(t) = p$ tenemos que $dp = f(t)dt$ y ya que $t = F^{-1}(F(t))$, llegamos a la siguiente definición de la curva de Lorenz

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)dt$$

donde $\mu = E(T)$. Usando esta forma de la curva de Lorenz, a partir de la ecuación (1.2) podemos verificar la equivalencia.

$$G = 1 - 2\int_0^1 L(p)dp = 1 - \frac{2}{\mu}\int_0^1\int_0^p F^{-1}(t)dtdp$$

Integramos por partes el segundo término de la igualdad, donde $u = \int_0^p F^{-1}(t)dt$, $v = p$, $du = F^{-1}(p)dp$ y $dv = dp$, entonces

$$\begin{aligned}
G &= 1 - \frac{2}{\mu}\left[p\int_0^p F^{-1}(t)dt\Big|_{p=0}^{p=1} - \int_0^1 pF^{-1}(p)dp\right] \\
&= 1 - \frac{2}{\mu}\left[\int_0^1 F^{-1}(t)dt - \int_0^1 pF^{-1}(p)dp\right] \\
&= 1 - \frac{2}{\mu}\left[\int_0^\infty xf(x)dx - \int_0^1 pF^{-1}(p)dp\right] \\
&= 1 - 2 + \frac{2}{\mu}\int_0^1 pF^{-1}(p)dp = \frac{1}{\mu}\int_0^1 2pF^{-1}(p)dp - 1 \\
&= \frac{1}{\mu}\int_0^\infty 2F(t)tf(t)dt - 1
\end{aligned}$$

Aquí usamos la sustitución $t = F(x)$ y $dt = f(x)dx$

Ahora sustituimos $p = F(t)$ y $dp = f(t)dt$

obtenemos el coeficiente de Gini en la forma

$$G = \frac{\int_0^\infty 2F(t)tf(t)dt}{\mu} - 1 \tag{1.3}$$

note que $\int_0^\infty 2F(t)tf(t)dt$ es $E(2F(T)T)$, así que podemos estimar el coeficiente de Gini por medio de promedios con los valores muestrales. Es decir, a través del estimador propuesto por Qin, Rao y Wu 2010

$$G_n = \frac{\frac{1}{n}\sum_{i=1}^{n} 2F_n(T_i)T_i}{\frac{1}{n}\sum_{i=1}^{n} T_i} - 1 \tag{1.4}$$

siendo $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq t)$ la función de distribución acumulada empírica de T, donde I es la función indicadora.

1.2. Procesos asintóticos

1.2.1. Secuencia

Una muestra aleatoria de tamaño n se define como un conjunto de variables aleatorias, $T_1, T_2, \ldots, T_n$, donde cada T_i representa el valor de una observación o experimentación aleatoria hecha en una población con función de densidad de probabilidad $f(t)$. Este conjunto de variables se caracteriza por tener una función de densidad de probabilidad conjunta $f(t_1, t_2, \ldots, t_n) = f(t_1)f(t_2)\cdots f(t_n)$.(Bain y Engelhardt 1992, p.159)

La población de la que se observa T_i se supone lo suficientemente grande (conceptualmente infinita) por lo que a pesar de extraer n individuos, con $n < \infty$, la distribución de probabilidad de los elementos restantes queda inalterada. Además, se asume que las observaciones se hacen de tal modo que las distribuciones de probabilidad de las demás observaciones no se ven afectadas. Por estas características estas variables también se consideran como un conjunto de **variables independientes e idénticamente distribuidas (v. i.i.d)**.

Las funciones de variables aleatorias son ellas mismas variables aleatorias, denotemos a estas últimas de la siguiente manera: $X_n = u(T_1, T_2, \ldots, T_n)$, es decir que X_n es una función de n variables aleatorias. En este sentido llamaremos **secuencia de variables aleatorias** a $X_1, X_2, \ldots$ con $X_1 = u(T_1), X_2 = u(T_1, T_2), \ldots, X_n = u(T_1, T_2, \ldots, T_n)$.

1.2.2. Convergencia

Definición. Sea X_n una secuencia de variables aleatorias tal que $X_n \sim H_n(x)$. Si $\lim_{n\to\infty} H_n(x) = H(x)$ se dice que X_n converge en distribución a X, ($X_n \xrightarrow{d} X$), donde $X \sim H(x)$, cuando esto ocurre H es la distribución asintótica de H_n.

Definición. Una secuencia X_n converge en probabilidad a (X, $X_n \xrightarrow{p} X$), si se cumple que $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$ para cualquier $\epsilon > 0$ arbitrariamente pequeña.

La convergencia en probabilidad implica la convergencia en distribución. Si $X_n \xrightarrow{p} \mu$, con μ constante, entonces $g(X_n) \xrightarrow{p} g(\mu)$ siempre que $g(\cdot)$ sea continua en μ. (Bain y Engelhardt 1992, p. 232, 247)

Un ejemplo de lo anterior es la **Ley de los grandes números** (LLN, por sus siglas en inglés) que dice que la secuencia de medias muestrales converge en probabilidad a la verdadera media poblacional, es decir $\overline{X}_n \xrightarrow{p} \mu$, siempre que la distribución tenga media y varianza finitas.

Otro ejemplo es la llamada *Ley Fuerte de los grandes números*, la cual afirma que dada la esperanza de una función $E(g(X)) = \int_{\infty}^{\infty} g(x)f(x)dx$, el estimador de la forma $g_n = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ converge casi seguramente a dicha esperanza, lo que se define como $P(\lim_{n\to\infty} g_n = E(g(X))) = 1$.

Los siguientes son algunos resultados principales de la convergencia de secuencias de variables aleatorias (Bain y Engelhardt 1992, p.248). Si X_n y Y_n son dos secuencias que convergen en probabilidad a dos constantes: $X_n \xrightarrow{p} a$ y $Y_n \xrightarrow{p} b$

1. $cX_n + dY_n \xrightarrow{p} ca + db$.
2. $X_nY_n \xrightarrow{p} ab$.
3. $X_n/a \xrightarrow{p} 1$, cuando $a \neq 0$.
4. $1/X_n \xrightarrow{p} 1/a$ si $P(X_n = 0) = 0$ para toda n.

Y el conocido teorema de Slutsky: si $X_n \xrightarrow{p} a$ y $Y_n \xrightarrow{d} Y$, sucede que

1. $X_n + Y_n \xrightarrow{d} a + Y$.
2. $X_nY_n \xrightarrow{d} aY$.
3. $Y_n/X_n \xrightarrow{d} Y/a$, cuando $a \neq 0$.

Una forma práctica de expresar la convergencia de las secuencias es con las notaciones $o_p(\cdot)$ y $O_p(\cdot)$. La primera, $o_p(\cdot)$, indica que una secuencia $X_n \xrightarrow{p} 0$ y se escribe

como $X_n = o_p(1)$, en general, para una secuencia positiva real que tiende a cero $r_n \to 0$ (por ejemplo $1/n$), si $r_n^{-1} X_n = o_p(1)$ entonces $X_n = o_p(r_n)$. (Kowalski y Tu 2008, sección 1.6.4)

$O_p(\cdot)$ indica que una secuencia tiene límites estocásticos, que se definen como sigue. Para cualquier $\epsilon > 0$ existe alguna M y N tal que $P(|X_n| \geq M) \leq \epsilon$ para toda $n \geq N$, y escribimos $X_n = O_p(1)$ o $X_n = O_p(r_n)$ para $r_n^{-1} X_n = O_p(1)$.

De forma más general, teniendo $\mathbf{X_n} = (X_{n1}, \ldots, X_{nl})^T \in \mathbb{R}^l$ para $l \geq 1$, utilizamos $\mathbf{X_n} = \mathbf{o}_p(1)$ cuando $\mathbf{X_n} \xrightarrow{p} \mathbf{0}$, y $\mathbf{X_n} = \mathbf{O}_p(1)$ cuando $\mathbf{X_n} \xrightarrow{d} \mathbf{X}$. Las principales propiedades y relaciones en esta notación, son (Kowalski y Tu 2008, p. 50):

1. $\mathbf{O}_p(1) + \mathbf{O}_p(1) = \mathbf{O}_p(1)$.
2. $\mathbf{O}_p(1) + \mathbf{o}_p(1) = \mathbf{O}_p(1)$.
3. $\mathbf{o}_p(1) + \mathbf{o}_p(1) = \mathbf{o}_p(1)$.
4. $\mathbf{O}_p(1)\mathbf{o}_p(1) = \mathbf{o}_p(1)$.
5. $\mathbf{O}_p(r_n) = \mathbf{o}_p(1)$.
6. $\mathbf{X_n} = \mathbf{o}_p(1)$ si, y sólo si, $X_{ni} = o_p(1)$ para toda $i = 1, \ldots, l$.
7. $\mathbf{X_n} = \mathbf{O}_p(1)$ si, y sólo si, $X_{ni} = O_p(1)$ para toda $i = 1, \ldots, l$.
8. Si $\mathbf{X_n} = \mathbf{o}_p(r_n)$ y $\mathbf{Y_n} = \mathbf{o}_p(q_n)$, se cumple que $\mathbf{X_n} + \mathbf{Y_n} = \mathbf{o}_p(max\{q_n, r_n\})$ y que $\mathbf{X_n^T Y_n} = \sum_{i=1}^{l} X_{ni} Y_{ni} = \mathbf{o}_p(q_n r_n)$.

1.2.3. Teorema central del límite

El teorema central del límite nos dice que dada una muestra aleatoria, T_i, cuya distribución tenga media μ y varianza σ^2 finita, la distribución asintótica de la estandarización de una suma de dichas variables

$$Z_n = \frac{\sum_{i=1}^{n} T_i - n\mu}{\sqrt{n\sigma^2}}$$

es normal estándar, es decir $Z_n \xrightarrow{d} Z$ o $N(0,1)$ cuando n tiende a infinito, y esto aunque las T_i no provengan necesariamente de una distribución normal ni sea necesariamente continua.

Además, para la secuencia X_n, si

$$Z_n = \frac{X_n - m}{v/\sqrt{n}} \xrightarrow{d} Z \sim N(0,1)$$

cuando n tiende a infinito, entonces X_n es una secuencia con distribución normal asintótica cuyas constantes m y $v/\sqrt{n}$ son su media asintótica y varianza asintótica, respectivamente, cumpliéndose al mismo tiempo que $X_n \xrightarrow{p} m$. (Bain y Engelhardt 1992, p. 238, 243, 246)

1.3. Análisis de supervivencia

El análisis de supervivencia (Survival Analysis en inglés) se describe, generalmente, como una colección de procedimientos estadísticos para modelar y analizar datos en los que la variable de interés es, si un evento ocurre, el tiempo que toma en ocurrir.

El tiempo se refiere a los segundos, horas, días, meses, etc., que pasan desde el seguimiento del fenómeno hasta que el evento de interés ocurre; también puede referirse a la "edad" de un individuo al momento de ocurrencia del evento. El evento puede ser una muerte, recuperación de una enfermedad, terminar la escuela, conseguir un trabajo, etc.(Kleinbaum y M. Klein 2012)

En el análisis de supervivencia, es usual referirse a una variable de tiempo como tiempo de supervivencia, como si fuera el tiempo que un individuo, o sujeto en el estudio, ha "sobrevivido" al evento que se espera. Al evento se le suele llamar falla, ya que este tipo de análisis se aplica principalmente en la medicina a cosas consideradas negativas, como una muerte, una enfermedad, etc.

1.3.1. Terminología

Denotaremos con una T mayúscula la variable aleatoria del tiempo de supervivencia (duración). Ya que T denota al tiempo, T puede ser cualquier número igual o mayor a 0. Así mismo, representaremos con una t minúscula cualquier valor específico de la variable aleatoria T.

Existen tres términos cuantitativos considerados en cualquier análisis de supervivencia: la función de supervivencia, denotada por $S(t)$, la función de riesgo o tasa de riesgo, denotada por $h(t)$ y la función de densidad de probabilidad, denotada como $f(t)$. Si conocemos alguna de las tres, entonces podemos determinar de manera única las restantes.

Función de Supervivencia

La función de supervivencia $S(t)$ es la medida básica para describir procesos en el análisis de supervivencia y da la probabilidad de que un sujeto sobreviva más de un tiempo t a la falla (ocurrencia del evento de interés), se define como

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

Si T es continua entonces $S(t)$ es también continua. $S(t)$ es también el complemento de la Función de Distribución Acumulada (FDA), es decir, $S(t) = 1 - F(t)$, donde $F(t) = P(T \leq t)$ o $S(t) = \int_t^{\infty} f(u)\ du$. Por lo que,

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

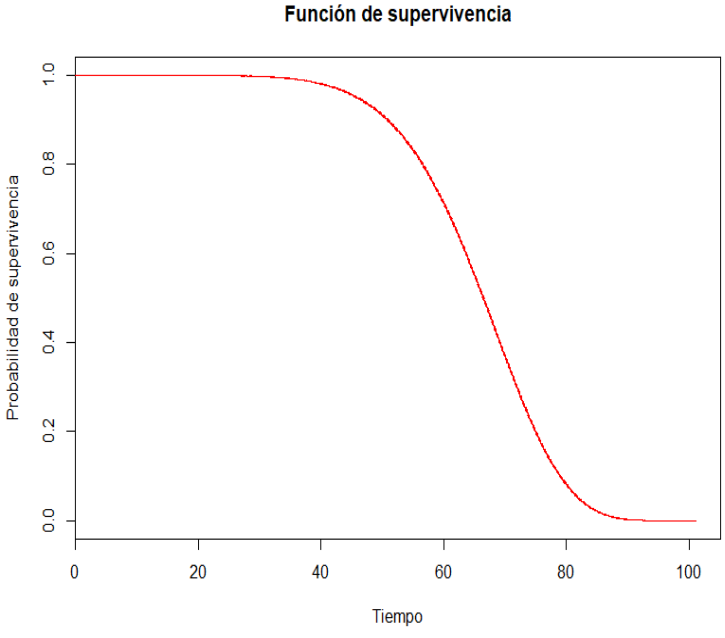

Figura 1.2: Ejemplo de la Función de supervivencia, color rojo.

Por ejemplo, si $T \sim Weibull(\lambda, \alpha)$, con FDP $f(t) = \alpha\lambda t^{\alpha-1} exp(-\lambda t^{\alpha})$ entonces $S(x) = e^{-\lambda t^{a}}$. Las funciones de supervivencia son siempre monótonas, no crecientes,

tales que $S(0) = 1$ y $S(\infty)$=0, iguales a 0 a medida que t tiende a infinito.

Función de Riesgo

La función de riesgo, $h\,(t)$, es dada por la fórmula

$$h\,(t) = \lim_{\Delta t \to 0} \frac{P\,(t < T \leq t + \Delta \mathrm{t} | T > t)}{\Delta t}$$

$h\,(t)$ es igual al límite cuando Δt se aproxima a cero.

La función de riesgo es una tasa de cambio de probabilidad instantánea por unidad de tiempo de que ocurra un evento, dado que el sujeto ha "sobrevivido" hasta el momento t. El valor de la función de riesgo, $h\,(t)$, sólo tiene la restricción $h(t) \geq 0$ así que puede exceder a 1 dependiendo de la unidad de medida (segundos, horas, días, etc.).

La función de riesgo también se puede representar de la forma equivalente

$$h\,(t) = f\,(t)\,/S\,(t) = -\frac{dS\,(t)\,/dt}{S\,(t)} = -dln\,[S\,(t)]\,/dt$$

Existe una relación entre la función de riesgo y la función de supervivencia, por lo que si se conoce una de las funciones se puede derivar la otra. La relación entre ambas funciones se muestra en las siguientes expresiones

$$H\,(t) = \int_0^t h\,(u)\,du = -ln\,[S\,(t)]$$

y

$$S\,(t) = exp\,[-H\,(t)] = exp\left[-\int_0^t h\,(u)\,du\right]$$

Si por ejemplo, $T \sim exp(\lambda)$, con FDP $f(t) = \lambda e^{\lambda t}$, entonces $h(t) = \lambda$. Mientras que las curvas de la función de supervivencia se parecen todas en su forma, las curvas de la función de riego pueden ser muy variadas, es por eso que se considera como más "informativa" que la función de supervivencia acerca del proceso de falla.

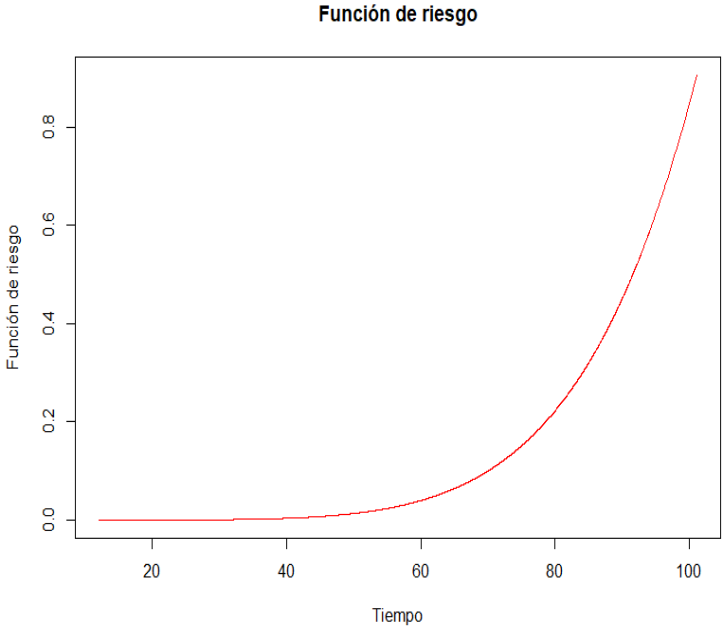

Figura 1.3: Ejemplo de la Función de riesgo, color rojo. Esta curva corresponde a una muestra obtenida de una $T \sim Weibull\,(1/70, 7)$.

Vida Residual Media

Este es otro parámetro importante en el análisis de supervivencia e indica, para el tiempo t, el tiempo medio restante antes de la ocurrencia del evento de interés. Se define como $m(t) = E(T - t|T > t)$, es decir

$$m(t) = E(T - t|T > t) = \frac{E((T-t)I_{(T>t)})}{P(T > t)} = \frac{\int_t^\infty (u-t)f(u)du}{S(t)}$$

donde si usamos el hecho de que $f(u)du = -dS(u)$, integrando por partes

$$m(t) = \frac{\int_t^\infty (u-t)f(u)du}{S(t)} = \frac{-(u-t)S(t)|_t^\infty + \int_t^\infty S(u)du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)}$$

ya que $S(\infty) = 0$. También, note que

$$m(0) = \frac{\int_0^\infty uf(u)du}{S(0)} = \frac{\int_0^\infty S(u)du}{1} = E(T) = \mu.$$

1.3.2. Censura

Dentro del análisis de supervivencia existe un problema clave que debe ser tomado en cuenta en todos los análisis, la censura. La censura es un problema de información incompleta (Guo 2010), en el que el investigador (observador) no puede observar el desarrollo completo hasta el evento de interés para todos los sujetos de estudio y es incapaz de determinar el momento exacto de la ocurrencia del evento para algunos sujetos del estudio.

Para el investigador, algunos de los sujetos de estudio están "censurados" hasta un valor de tiempo, por lo que a algunos de ellos les ocurrirá el evento en el futuro, a otros tal vez nunca y a otros más podría ocurrirles un evento diferente al de interés que los excluya (riesgo competitivo). Ignorar las observaciones censuradas provoca pérdida de información importante y sesgo en los resultados, haciéndolos poco confiables.

En los casos en que los sujetos son censurados porque el tiempo de seguimiento terminó, porque el sujeto se "extravió" (se perdió a la observación del investigador), porque se retiró o porque le ocurrió un evento diferente al esperado, sabemos que el periodo es incompleto por la derecha. Llamamos a este tipo de datos censurados por la derecha. Los datos pueden también ser censurados por la izquierda o pertenecer a un intervalo censurado.

Censura por la derecha

La censura por la derecha incluye las clasificaciones: *Censura Tipo I*, *Censura Tipo II* y *Censura por Riesgos Competitivos*.

La Censura Tipo I es aquella donde el evento es observado solo si ocurre antes de un momento fijado por el investigador, estos tiempos de censura pueden variar de un elemento a otro, como veremos, cuando los momentos de entrada son diferentes o existen diferentes "cortes" en el estudio.

Por diferentes causas, es posible que el observador tenga que terminar su estudio o realizar las mediciones antes de que todos los individuos hayan experimentado el evento de interés. Si los individuos no se "perdieron" o se retiraron del estudio, las observaciones censuradas tienen un tiempo de censura igual al término del estudio. A pesar de que el tiempo de término del estudio es fijado por el investigador, algunos tiempos de censura son fijos y otros aleatorios porque las unidades muestrales pueden entrar en diferentes tiempos. (J. P. Klein y Moeschberger 2003).

Asumiremos que los sujetos del estudio tienen un tiempo de falla, T, y un tiempo de censura, C, donde las variables T son independientes e idénticamente distribuidas de acuerdo con la función de densidad de probabilidad $f(t)$ con función de supervivencia $S(t)$.

El tiempo T será observable sólo cuando, para un elemento dado, $T < C$. Si $T > C$, el individuo ha sobrevivido y su tiempo de falla se censura en C. Resumidamente, podemos representar la información del estudio con el par de variables aleatorias (X, δ), donde $X = min(T, C)$ mientras que $\delta = 1$ si $X = T$ o $\delta = 0$ si $X = C$.

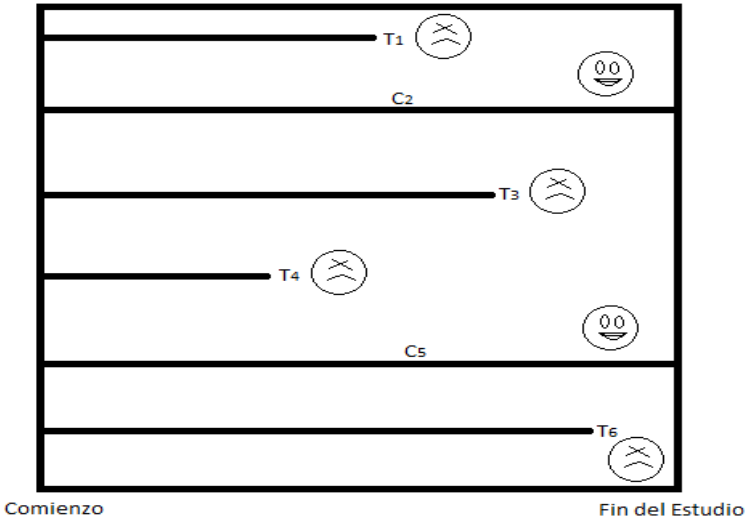

Figura 1.4: Ejemplo de censura de Tipo I básica.

Llamamos *censura generalizada de Tipo I* a la forma de censura que no restringe la entrada de los individuos al estudio en un sólo momento, es decir cuando los individuos entran en diferentes tiempos, determinando diferentes valores de C por este motivo a pesar de que el estudio finalice al mismo tiempo para todos.

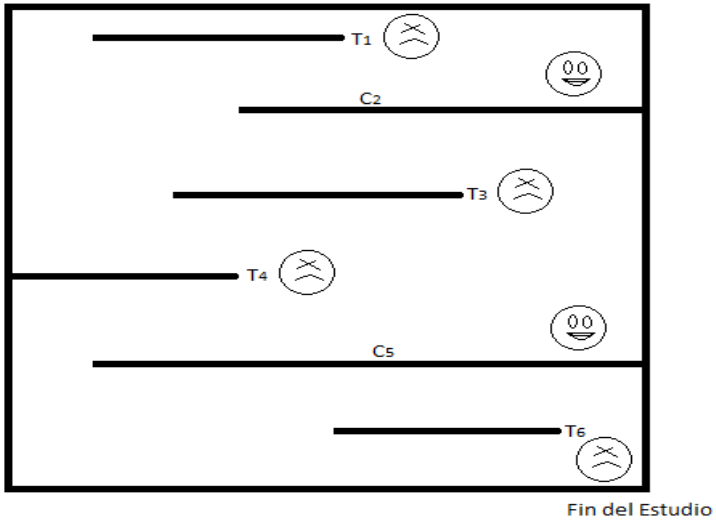

Figura 1.5: Ejemplo de censura generalizada de Tipo I.

Además, puede darse el caso de que exista más de un momento de medición fijado por el investigador que provoque distintos momentos de censura. Por ejemplo, si un estudio dispone en un inicio de un conjunto grande de individuos hasta un primer momento dado, en el que una parte de los individuos tendrá que retirarse, continuando el estudio con la parte restante de los sujetos, en ese primer momento habrá de darse una medición que determine cuantas fallas hubieron y sus tiempos, pero también habrán de registrarse tiempos censurados para aquellos individuos sobrevivientes de ese primer conjunto. Nos referimos a la *censura progresiva de Tipo I.*

El segundo tipo de censura, la *Censura de Tipo II*, consiste en detener el estudio cuando se registran las primeras r fallas, donde $(r < n)$. Todos los individuos inician juntos el estudio, el número de fallos y elementos censurados es fijo, sin embargo el tiempo de censura C para los $(n - r)$ sujetos restantes es una magnitud aleatoria que depende de los r primeros fallos.

El tercer tipo de censura por la derecha es la *Censura por Riesgos Competitivos*, cuyo caso especial es la censura aleatoria. Un riesgo competitivo puede definirse como la posibilidad de ocurrencia de un evento que sea mutuamente excluyente con el evento de interés.J. P. Klein y Moeschberger 2003, p.69

Pensemos en un estudio en el que estamos interesados en la distribución marginal

de los tiempos en que sucede un evento de interés, pero en el que los individuos del estudio están expuestos a sufrir una "falla" distinta a la que interesa, que los "retire" en el sentido de que haga imposible observar el evento de interés. Dichos individuos quedarían censurados aleatoriamente por la derecha.

Por ejemplo, una persona desempleada de la que se espera consiga trabajo, pero en lugar de ello sale del país y no se le puede dar seguimiento, o comienza una carrera universitaria que le impida trabajar, estos son riesgos competitivos.

Para ser capaces de identificar las distribuciones marginales de los riesgos competitivos, por ejemplo, de nuestro evento de interés, debemos suponer que los tiempos de censura y los tiempos del evento de interés sean independientes entre sí. Ejemplos de los que se pueden considerar como tiempos censurados independientes del evento de interés son migraciones, muertes, retiro voluntario inesperado del estudio, etc.

Verosimilitud

En el análisis de supervivencia se usa la función de verosimilitud como base del análisis inferencial, tanto en modelos paramétricos como en modelos semiparamétricos.

Dado un proceso de censura aleatoria, siendo T y C variables aleatorias independientes, donde cada elemento i tiene un tiempo hasta el evento de interés (*tiempo de vida*) T_i y un tiempo de censura C_i. C tiene como función de densidad $f_{_C}(c)$ y, $S_{_C}(c)$, como función de supervivencia y, además, $f(t,c)$ como función de densidad conjunta. Como antes, $X_i = min(T_i, C_i)$ y $\delta_i = 1$ si $X = T$ o $\delta_i = 0$ si $X = C$. Una muestra consiste en el conjunto de pares (X_i, δ_i), con $i = 1 \ldots n$ y su función de densidad conjunta se puede obtener a partir de $f(t,c)$

$$\begin{aligned} P(X_i \leq x, \delta = 1) &= P(T_i \leq x, C_i > T_i) \\ &= \int_0^x \int_u^\infty f(u,v)dvdu. \end{aligned}$$

Por lo que

$$f(x,1) = \frac{\partial}{\partial x} \int_0^x \int_u^\infty f(u,v)dvdu.$$

Si T y C son independientes, entonces

$$= \int_0^x \int_u^\infty f_t(u) f_{_C}(v) dv du$$
$$\Psi_1(x) = \int_0^x f_t(u) S_{_C}(u) du = F(x, \delta = 1).$$

Así que

$$f(x,1) = \frac{\partial \Psi_1(x)}{\partial x} = f_t(x) S_{_C}(x). \tag{1.5}$$

De la misma manera

$$\begin{aligned} P(X_i \leq x, \delta = 0) &= P(C_i \leq x, T_i > C_i) \\ &= \int_0^x \int_v^\infty f(u,v) du dv \\ &= \int_0^x \int_v^\infty f_t(u) f_{_C}(v) du dv \\ \Psi_0(x) &= \int_0^x S_t(v) f_{_C}(v) dv = F(x, \delta = 0), \text{ para el caso de independencia.} \end{aligned}$$

Por lo que

$$f(x,0) = \frac{\partial \Psi_0(x)}{\partial x} = S_t(x) f_{_C}(x). \tag{1.6}$$

Así que la verosimilitud la podemos expresar en una sola ecuación como

$$L = \prod_{i=1}^n \left(f(x_i) S_{_C}(x_i) \right)^{\delta_i} \left(f_{_C}(x_i) S_t(x_i) \right)^{1-\delta_i}.$$

En caso de que T y C no fueran independientes, su verosimilitud sería de la siguiente forma

$$L \propto \prod_{i=1}^n \left([-\partial S(t, x_i)/\partial x]_{t=x_i} \right)^{\delta_i} \left([-\partial S(x_i, c)/\partial c]_{c=x_i} \right)^{1-\delta_i}$$

donde $S(t,c)$ es la función de supervivencia conjunta.

1.3.3. Proceso de conteo y martingalas

Debemos comenzar con algunos conceptos y definiciones de la teoría de la probabilidad, que están expuestos de manera concisa en Fleming y Harrington 1991, sobre todo del Apéndice A y el capítulo 1, de donde nos basamos.

Dado un conjunto no vacío Ω formado por elementos ω y una colección de subconjuntos $\mathcal{A}$ de Ω, se dice que $\mathcal{A}$ es una σ-álgebra si el subconjunto $\bar{E}$ (complemento de E) pertenece a $\mathcal{A}$ siempre que $E \in \mathcal{A}$, y $\cup_{i=1}^{\infty} E_i \in \mathcal{A}$ cuando $E_i \in \mathcal{A}$, para $1 \leq i < \infty$.

$(\Omega, \mathcal{F}, P)$ es llamado espacio de probabilidad si $\mathcal{F}$ una σ-álgebra sobre Ω y P es una medida de probabilidad definida en $\mathcal{F}$. Dado un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, una variable aleatoria $X : \Omega \to \mathbb{R}$ en una función de valor real de Ω a $\mathbb{R}$ que satisface $\{\omega : -\infty < X(\omega) \leq b\}$, con $\omega \in \mathcal{F}$ por ello se dice que X es *medible* con respecto a $\mathcal{F}$ o que X es $\mathcal{F}$-medible.

Un proceso estocástico es una familia de variables aleatorias $X = \{X(t) : t \in \Gamma\}$ definidas en el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$. El conjunto Γ denota el tiempo, que puede ser discreto o continuo. Del proceso estocástico X, las funciones aleatorias $X(\cdot, \omega)$ son los *pasos muestrales* o *trayectorias muestrales* de X. El proceso será de variación acotada, creciente, continuo por la izquierda, continuo por la derecha, con límites por la derecha o por la izquierda si el conjunto de las trayectorias muestrales con alguna de las propiedades anteriores tiene probabilidad 1.

Una *filtración* $\{\mathcal{F} : t \geq 0\}$ de un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ es una familia de σ-álgebras tal que $\mathcal{F}_t \subset \mathcal{F}$ para toda t, y $\mathcal{F}_s \subset \mathcal{F}_t$ para $s \leq t$. Las filtraciones son continuas por la derecha si $\mathcal{F}_{t+} = \mathcal{F}_t$, estas son las filtraciones que se deben asumir para procesos estocásticos escalonados, continuos por la derecha.

Se dice de un proceso $X(t)$ que es adaptado a la filtración $\mathcal{F}_t$ si es $\mathcal{F}_t$-medible para toda t. Sea $\mathcal{F}_t = \sigma\{X(s) : 0 \leq s \leq t\}$, la σ-álgebra más pequeña que hace medibles a todas las variables aleatorias $X(s)$, entonces la filtración $\mathcal{F}_t$ es también llamada la historia de X.

Un proceso de conteo $N(t), t \geq 0$, es un proceso estocástico adaptado a la filtración $\mathcal{F}_t$, tal que $N(0) = 0$ y $N(t) < \infty$ para t finito, con probabilidad 1. Cuando hay censura por la derecha, el proceso se define por $N_i(x) = I(X_i \leq x, \delta_i = 1)$, es decir con la función indicadora por lo que $N_i = 1$ solo cuando el tiempo observado (no censurado pues $\delta_i = 1$) es igual o menor a x; $N(x) = \sum_{i=1}^{n} N_i(x)$ también es un proceso de conteo que señala el número de fallas hasta el tiempo x. En este trabajo utiliza-

remos también el proceso de conteo $N_i^C = I(X_i \leq x, \delta_i = 0)$ y $N^C(x) = \sum_{i=1}^{n} N_i^C(x)$ que se realizan desde el punto de vista de C, la variable de censura. La filtración correspondiente será $\mathcal{F}_x = \sigma\{T_i, I(X_i \leq u, \delta_i = 0), 0 \leq u \leq x, i = 1, \ldots, n\}$, generada por el tiempo de fallo y el proceso $N_i^C(u)$, para $u \leq x$.

En un proceso de conteo con datos censurados por la derecha podemos definir $dN(x) = N[(x+dx)^-] - N(x-)$, es decir el cambio en $N(x)$ dado un cambio pequeño del tiempo x.[2] Por lo general, dN será igual a 0 cuando no sucedan fallas en el tiempo x e igual a 1 si una falla ocurre en x.

Definiremos también $R(x) = \sum_{i=1}^{n} R_i(x)$, donde $R_i(x) = I(X_i \geq x)$, como el proceso de riesgo que es la cantidad de individuos que han sobrevivido al menos hasta el tiempo x; $\phi(x) = R(x)h(x)$ como el proceso de intensidad, que es también un proceso estocástico; $\Phi(x) = \int_0^x \phi(u)du$, con $x \geq 0$, como el proceso de intensidad acumulado y; $M(x) = N(x) - \Phi(x)$ como el *proceso de conteo martingala*, porque de hecho es una martingala, lo que explicaremos a continuación.

Martingalas

Una martingala es un proceso estocástico que se caracteriza porque su esperanza al tiempo x, dada la historia $\mathcal{F}_y$ hasta el tiempo $y < x$, es igual a su valor en y. En otras palabras, $M(x)$ es una martingala si: $E(M(x)|\mathcal{F}_y) = M(y)$ para cualquier $y < x$, Las martingalas se consideran constantes dada una historia de ellas cumpliéndose por ello también que $E(M(y)|\mathcal{F}_x) = M(y)$ con $y < x$ como antes.

Esta característica de las martingalas implica varias propiedades que las hacen de relativo fácil manejo. Para empezar, asumiendo que $M(0) = 0$, por el hecho de que $E(M(x)) = E(E(M(x)|\mathcal{F}_0)) = 0$, decimos que la martingala tiene media 0 para cualquier tiempo x en un intervalo finito $(0, \mathcal{X}]$.

La esperanza de un incremento dada una historia es iguala 0, es decir $E(dM(x)|\mathcal{F}_{x-}) = E(M(x) - M(x-)|\mathcal{F}_{x-}) = E(M(x)|\mathcal{F}_{x-}) - M(x-) = 0$. La

$$\begin{aligned} cov(dM(u), dM(v)) =& E\big[\big(M(u) - M(u-)\big)\big(M(v) - M(v-)\big)\big] \\ =& E\big[E\big((M(u) - M(u-))(M(v) - M(v-))|\mathcal{F}_u\big)\big] \\ =& E\big[\big(M(u) - M(u-)\big)E\big(M(v) - M(v-)|\mathcal{F}_u\big)\big] \\ =& E\big[\big(M(u) - M(u-)\big)\big(0\big)\big] = 0, \end{aligned}$$

[2] La notación x^- indica el tiempo justo antes de x.

para $u- < u < v- < v$. Los incrementos no están correlacionados.

Hay dos procesos que explican la variación de una martingala. El primero es el *proceso de variación predecible*, denotado por $\langle M \rangle$ y definido por

$$\langle M \rangle(x) = \lim_{n \to \infty} \sum_{i=1}^{n} Var(\Delta M_i | \mathcal{F}_{(i-1)x/n}),$$

el segundo, llamado *proceso de variación opcional*, denotado por $[M]$ y definido por

$$[M](x) = \lim_{n \to \infty} \sum_{i=1}^{n} (\Delta M_i)^2.$$

Para ambos, el intervalo $[0, x]$ se divide en n subintervalos de tamaño x/n y $\Delta M_i = M(ix/n) - M((i-1)x/n)$. Derivado del proceso de variación predecible, $d\langle M \rangle(x) = Var(dM(x)|\mathcal{F}_{x-})$.(O. Aalen, Borgan y Gjessing 2008, sección 2.2.1)

De manera análoga se definen los procesos de covariación predecible y opcional entre dos martingalas M_1 y M_2, $\langle M_1, M_2 \rangle(x)$ y $[\langle M_1, M_2]](x)$, pero sustituyendo $Var(\Delta M_i | \mathcal{F}_{(i-1)x/n})$ con $Cov(\Delta M_{1i}, \Delta M_{2i} | \mathcal{F}_{(i-1)x/n})$ y $(\Delta M_i)^2$ con $(\Delta M_{1i} \Delta M_{2i})$, respectivamente.

Un resultado importante del proceso de variación predecible para este trabajo es el siguiente: $M^2(x) - \langle M \rangle(x)$ es también una martingala de media 0 y, en consecuencia, $Var(M(x)) = E(M^2(x)) = E(M^2(x) + \langle M \rangle(x) - \langle M \rangle(x)) = E[\langle M \rangle(x)]$.

Por su parte, $M_1 M_2 - \langle M_1, M_2 \rangle$ forman una martingala de media 0 y $d\langle M_1, M_2 \rangle(x) = cov(dM_1(x), dM_2(x) | \mathcal{F}_{x-})$. En tiempo continuo dos procesos de conteo N_1 y N_2 no tiene saltos al mismo tiempo, así que sus martingalas correspondientes M_1 y M_2 no están correlacionadas: $\langle M_1, M_2 \rangle(x) = 0$ y se dice que las martingalas M_1 y M_2 son ortogonales. Puesto que $M_1 M_2 - \langle M_1, M_2 \rangle$ son una martingala, $M_1 M_2$ es a su vez una martingala. (O. Aalen, Borgan y Gjessing 2008, p. 50, 56)

Integral Estocástica

Un proceso estocástico predecible, $\mathcal{H}(x)$, es un proceso estocástico cuyo valor es conocido hasta justo antes de x, lo que se denota por $E(\mathcal{H}(x)|\mathcal{F}_{x-}) = \mathcal{H}(x)$. Se explica con la historia $\mathcal{F}_x$ (es $\mathcal{F}_x$-medible), $\mathcal{H}(x)$ es una función de pasos muestrales continuos por la izquierda.

Entonces, una integral estocástica está definida como

$$I(x) = \int_0^x \mathcal{H}(u)dM(u),$$

y es ella misma una martingala de media 0 con respecto a $\mathcal{F}$. Su proceso de variación predecible es

$$\Big\langle \int \mathcal{H}(u)dM(u) \Big\rangle = \int \mathcal{H}(u)^2 d\langle M\rangle(u).$$

En el contexto de los procesos de conteo, $N(x)$ es una submartingala. Una submartingala es cualquier proceso no decreciente que por lo general aumenta con el paso del tiempo, satisface que $E(N(x)|\mathcal{F}_y) \geq N(y)$ con $y < x$. El proceso de intensidad, $\phi(x) = R(x)h(x)$, es un proceso predecible que puede definirse como: $\phi(x)dx = E(dN(x)|\mathcal{F}_{x-}) = P\big(dN(x) = 1|\mathcal{F}_{x-}\big)$. El proceso de intensidad acumulado $\Phi(x) = \int_0^x \phi(u)du$ es conocido como el compensador de $N(x)$ al formar, como señalamos arriba, una martingala $M(x) = N(X) - \int_0^x \phi(u)du$.

En este contexto, el proceso de variación predecible del diferencial de la martingala se define como: $d\langle M\rangle(x) = Var(dM(x)|\mathcal{F}_{x-}) = Var[dN(x) - \phi(x)dx|\mathcal{F}_{x-}] = Var[dN(x)|\mathcal{F}_{x-}]$, ya que $\phi(x)$ es predecible, es decir constante dada la historia $\mathcal{F}_{x-}$, lo que lleva a

$$\begin{aligned} d\langle M\rangle(x) =& E[(dN(x) - \phi(x)dx)^2|\mathcal{F}_{x-}] \\ =& E[dN(x)^2 - 2dN(x)\phi(x)dx + (\phi(x)dx)^2|\mathcal{F}_{x-}] \\ \approx& E[dN(x)^2|\mathcal{F}_{x-}] - (\phi(x)dx)^2 \\ =& E[dN(x)|\mathcal{F}_{x-}] - (\phi(x)dx)^2 \approx \phi(x)dx, \end{aligned}$$

porque $(\phi(x)dx)^2$ es muy cercano a 0. De aquí el siguiente resultado

$$\langle M\rangle(x) = \int_0^x \phi(u)du = \int_0^x R(u)h(u)du,$$

el proceso acumulado de intensidad es también el compensador de $M^2(x)$, recordemos que $M^2(x) - \langle M\rangle(x)$ es una martingala de media 0. (O. Aalen, Borgan y Gjessing 2008, sección 2.2.5)

La integral estocástica correspondiente a un proceso de conteo se define ahora como

$$I(x) = \int_0^x \mathcal{H}(u)dM(u) = \int_0^x \mathcal{H}(u)dN(u) - \int_0^x \mathcal{H}(u)R(u)h(u)du,$$

donde $\int_0^x \mathcal{H}(u)dN(u) = \sum_{X_i \leq x} \mathcal{H}(X_i)dN_i(X_i)$. Mientras que el proceso de variación predecible correspondiente es

$$\Big\langle \int \mathcal{H}(u)dM(u) \Big\rangle(x) = \int \mathcal{H}(u)^2 d\langle M\rangle(u) = \int \mathcal{H}(u)^2 R(u)h(u)du.$$

En este trabajo analizaremos martingalas locales, es decir que tienen la propiedad de ser martingalas para intervalos cerrados sobre la recta real $[0, x]$, $x \leq \infty$. Una propiedad de un proceso estocástico se mantiene localmente si dicha propiedad es cumplida por un proceso de paro $X_n = \{X(min(x, \tau_n)) : x \geq 0\}$, con τ_n como una secuencia creciente de tiempos aleatorios o *tiempos de paro*, tal que $\lim_{n\to\infty} \tau_n = \infty$, las variables aleatorias τ son tiempos de paro con respecto a $\mathcal{F}_x$ si $\{\tau \leq x\} \in \mathcal{F}$. Con estas dos características, una secuencia creciente de tiempos de paro $\tau_n = 1, \ldots, n$ se conoce como una *secuencia de localización*. (Fleming y Harrington 1991, Sección 2.2)

Así que, una martingala local respecto de $\mathcal{F}_x$ es un proceso estocástico tal que una secuencia de localización τ_n existe, de la que para cada n, $M_n = \{M(min(x, \tau_n)) : 0 \leq x < \infty\}$ es una $\mathcal{F}_x$-martingala. Un proceso adaptado $X(x)$ es localmente acotado si $X_n = \{X(min(x, \tau)) : x \geq 0\}$ es un proceso acotado para una secuencia de localización apropiada τ_n. Esta última propiedad la utilizaremos para los integrandos de las integrales estocásticas que surjan en el análisis. (Fleming y Harrington 1991, Definición 2.2.3)

Teorema central del límite para martingalas

$W(b)$ es el valor del *proceso de Wiener* en el tiempo b, dado un pequeño intervalo de tiempo $(a, b]$, la diferencia $W(b) - W(a)$ se distribuye normalmente con $E(W(b) - W(a))$ y $Var(W(b) - W(a)) = b - a$. El proceso de Wiener, o *movimiento browniano*, surge como límite en las integrales estocásticas de procesos de conteo martingala, el proceso de Wiener tiene pasos muestrales continuos y sus incrementos en intervalos que no se traslapan son independientes.

Sea $U^\infty(x)$ una *martingala Gaussiana* de media 0 y proceso de variación predecible $\langle U\rangle(x) = V(x)$, donde $V(x) = \int_0^x v(u)du$ es una función estrictamente creciente con $V(0) = 0$. Sucede que $U^\infty(x) = W(V(x))$ y $U^\infty(x)$ mantiene las propiedades del proceso de Wiener mencionadas arriba. (O. Aalen, Borgan y Gjessing 2008, sección 2.3.1)

A medida que una secuencia de procesos de conteo crece (n es más grande), el número de saltos aumenta y se hacen más densos, lo que en el límite converge a una martingala de pasos muestrales continuos. Cuando en el límite el proceso de variación predecible de una martingala converge a una función determinista $V(x)$, dicha martingala converge a una martingala Gaussiana.(O. Aalen, Borgan y Gjessing 2008,

sección 2.3.2)

A continuación, presentamos los resultados básicos del teorema del límite central para integrales estocásticas respecto de procesos de conteo martingala, los tomamos de Fleming y Harrington 1991, sección 5.1.

Sea

$$U^{(n)}(x) = \sum_{i=1}^{n} \int_0^x \mathcal{H}_i^{(n)}(u) dM_i^{(n)}(u)$$

una martingala (cada i-ésimo elemento lo es también). En el caso general hay r estadísticos $U_l^{(n)}$ para $l = 1, \ldots, r$, que suelen hacer referencia a las muestras involucradas (para este trabajo $r = 1$).

$$U_l^{(n)}(x) = \sum_{i=1}^{n} \int_0^x \mathcal{H}_{i,l}^{(n)}(u) dM_{i,l}^{(n)}(u)$$

con $M_{i,l}^{(n)}(x) = N_{i,l}^{(n)}(x) - \int_0^x \phi_{i,l}^{(n)}(u)du$ que es una martingala de cuadrado integrable local. Además, se define

$$U_{l,\epsilon}^{(n)}(x) = \sum_{i=1}^{n} \int_0^x \mathcal{H}_{i,l}^{(n)}(u) I(|\mathcal{H}_{nj}(u)| \geq \epsilon) dM_{i,l}^{(n)}(u)$$

que contiene solo los saltos del proceso $U_l^{(n)}(x)$ iguales o mayores a algún $\epsilon > 0$. Para l fija, a medida que $n \to \infty$, resulta que

$$\langle U^{(n)} \rangle(x) \xrightarrow{p} V(x), \tag{1.7}$$

donde $V(x)$ es la misma de arriba, y

$$\langle U_\epsilon^{(n)} \rangle(x) \xrightarrow{p} 0, \tag{1.8}$$

entonces $U^{(n)}(x) \xrightarrow{d} U^{(\infty)}(x)$.

En otras palabras, el límite central para martingalas dice que si se cumplen estas dos condiciones una secuencia de integrales estocásticas (o suma de ellas) converge a la martingala Gaussiana de media 0 que, para un valor dado x, se distribuye como una variable aleatoria normal, con media 0 y varianza $V(x)$.

1.3.4. Estimador Kaplan-Meier

El estimador Kaplan-Meier es una alternativa para calcular las probabilidades de supervivencia y para estimar la curva de supervivencia. Dados los tiempos en que se registra el evento (no censurados) $t_1 < t_2 < \ldots < t_k$, podemos definir el estimador KM de las siguientes maneras:

$$\begin{aligned}\widehat{K}(t_i) &= \widehat{K}\left(t_{(i-1)}\right) \cdot P\left(T > t_i | \, T \geq t_i\right) \\ &= \prod_{j=1}^{i} \widehat{P}\left(T > t_j | \, T \geq t_j\right)\end{aligned}$$

o también

$$\widehat{K}(t) = \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right] \tag{1.9}$$

para toda $t \geq t_1$, donde d_i es el número de fallas al tiempo t_i y Y_i es el número de sujetos con tiempo de supervivencia $t \geq t_i$. En caso de que $t < t_1$, entonces $\widehat{K}(t) = 1$.

Para los estudios que incluyen tiempos censurados es conveniente usar la siguiente forma del estimador Kaplan-Meier, ya que parte directamente de la variable X (recuerde que $X = min(T, C)$), y no hay necesidad de hacer un filtro previo de los tiempos de fallo pues se utilizan los procesos de conteo

$$\widehat{K}(u) = \prod_{x_i \leq u} \left[1 - \frac{dN(x_i)}{R(x_i)}\right] \tag{1.10}$$

De manera más general, cualquier fórmula del estimador Kaplan-Meier de una probabilidad de supervivencia consiste en el producto de los términos hasta el tiempo de supervivencia que se especifica. Es por eso que el estimador Kaplan-Meier se conoce también como una fórmula de "límite de producto".

Representación como Integral-Producto

Existe una forma muy útil de relacionar el estimador Kaplan-Meier con la función de riesgo. Recordemos que suponiendo que la función de supervivencia $S(t)$ es continua, se relaciona con la función de riesgo acumulada de la siguiente manera

$$S(t) = e^{-H(t)} = e^{-\int_0^t h(u)du} = e^{-\int_0^t -S'(u)/S(u)du}.$$

Pero hay otras formas más generales de relacionar estas funciones y sus estimadores que no necesitan que dichas funciones sean completamente continuas o completamente discretas. (O. Aalen, Borgan y Gjessing 2008, Sección 3.2.4)

Mientras $S(t)$ sea al menos continua por la derecha con límites por la izquierda[3], tendremos que dS(t) será un cambio pequeño en S debido a un cambio pequeño (dt) y $-dS(t) = P(t \leq T < t + dt)$, así que

$$h(t)dt = P(t \leq T < t + dt | T \geq t) = \frac{-dS(t)}{S(t-)} \quad \text{y} \quad H(t) = -\int_0^t \frac{dS(u)}{S(-u)}$$

En donde si $S(t)$ es completamente continua entonces $dS(u) = -f(u)du$ y $S(u-) = S(u)$.(O. Aalen, Borgan y Gjessing 2008, Apéndice A.1)

Una herramienta conceptual que nos permite expresar la función de supervivencia de cualquier distribución a través de la función de riesgo acumulada es la *integral-producto*. La integral-producto se basa en la partición del intervalo $[0, t]$ en un número grande de intervalos pequeños con $0 = t_0 < t_1 < \ldots < t_k = t$ y el correspondiente producto

$$\prod_{j=1}^{k} (1 - H(t_j) - H(t_{j-1})).$$

A medida que aumenta el número de intervalos y su longitud tiende a 0 el producto se aproxima a un límite en forma análoga a la integral que representa el límite de sumas finitas en un intervalo. Se denota por

$$S(t) = \pi_{u \leq t}(1 - dH(u)).$$

Cuando $S(t)$ es completamente continua, $\pi_{u \leq t}(1 - dH(u)) = e^{-H(t)}$ y si $S(t)$ es completamente discreta $\pi_{u \leq t}(1 - dH(u)) = \prod_{u_i \leq t} (1 - h(u_i))$. Y en el caso más general, pudiendo descomponer $H(t) = H_c(t) + H_d(t)$, en que H_c es la parte continua y H_d la parte discreta. Quedando $S(t)$ con la la integral-producto como

$$\pi_{u \leq t}(1 - dH(u)) = e^{-Hc(t)} \prod_{u_i \leq t} (1 - h_d(u_i)).$$

[3] A esto se le conoce como una función *Cadlag*, puede ver un ejemplo grafico en:
Càdlàg. En Wikipedia, The Free Encyclopedia . Consultado el 02:18, 12 de febrero de 2019, de https://en.wikipedia.org/wiki/Càdlàg.

Veamos que $\widehat{H}(t) = \sum_{T_i \leq t} d\widehat{H}(T_i) = \int_0^t d\widehat{H}(u)$[4], donde $d\widehat{H}(x) = \frac{dN(x)}{R(x)}$ para $Y(x) > 0$, mejor conocido como el estimador *Nelson-Aalen*. Este estimador se usó más arriba para definir el estimador Kaplan-Meier que no es más que

$$\widehat{K}(t) = \widehat{S}(t) = \prod_{u \leq t}(1 - d\widehat{H}(u)) = \prod_{T_i \leq t}(1 - d\widehat{H}(T_i)).$$

Propiedades

Partamos del cociente, $\frac{dM(x)}{R(x)} = \frac{dN(x)}{R(x)} - h(x)dx$, para evitar divisiones entre 0, introducimos la función indicadora $J(x) = I(R(x) > 0)$; esta notación lleva a la indeterminación cuando $R(x) = 0$, pero como se menciona en toda la literatura al respecto donde se utiliza la misma notación, se conviene que $\frac{J(x)}{R(x)} = 0$ cuando $R(x) = 0$. Tendremos ahora $\frac{J(x)}{R(x)}dM(x) = \frac{J(x)}{R(x)}dN(x) - J(x)h(x)dx$ que en forma de integral resulta

$$\int_0^x \frac{J(u)}{R(u)}dM(u) = \int_0^x \frac{J(u)}{R(u)}dN(u) - \int_0^x J(x)h(u)du.$$

Los términos de la diferencia del lado derecho son: el estimador Nelson-Aalen

$$\widehat{H}(x) = \int_0^x \frac{J(u)}{R(u)}dN(u)$$

y la función

$$H^*(x) = \int_0^x J(x)h(u)du$$

que asintóticamente es igual $H(x)$, la función acumulada de riesgo. Note ahora que $\int_0^x \frac{J(u)}{R(u)}dM(u)$ es una integral estocástica respecto de la martingala M, así que tiene media 0 y en consecuencia $E(\widehat{H}(x) - H^*(x)) = 0$, es decir que el estimador Nelson-Aalen es insesgado respecto de $H^*(x)$.

Vamos a aplicar el teorema central del límite para martingalas a la martingala

$$\sqrt{n}(\widehat{H}(x) - H^*(x)) = \int_0^x \sqrt{n}\frac{J(u)}{R(u)}dM(u).$$

Asumimos que $\frac{R(x)}{n} \xrightarrow{p} r(x)$ a medida que $n \to \infty$, para un tiempo dado la proporción de individuos en riesgo se vuelve estable. Con $M(x) = \sum_{i=1}^n M_i(x)$ y $R(x) =$

[4]Esta última forma es una integral de *Lebesgue-Stieltjes*.

$\sum_{i=1}^{n} R_i(x)$, (1.7) y (1.8) se cumplen ya que

$$\mathcal{H}(u)^2\phi(u) = \frac{J(u)h(u)}{R(u)/n} \xrightarrow{p} \frac{h(u)}{r(u)}$$

y

$$\mathcal{H}(u) = \frac{1}{\sqrt{n}}\frac{J(u)}{R(u)/n} \xrightarrow{p} 0,$$

de modo que, para x dado, $\sqrt{n}(\widehat{H}(x) - H^*(x))$ converge a una martingala Gaussiana con función de varianza $V(x) = \int_0^x \frac{h(u)}{r(u)}du$. (O. Aalen, Borgan y Gjessing 2008, sección 3.1.6)

Por otro lado, utilizando las expresiones de la función de supervivencia

$$S^*(x) = \prod_{s\in(0,x]}(1 - dH^*(s)),$$

que análogamente es asintóticamente igual a $S(x)$, y del estimador Kaplan-Meier

$$\widehat{K}(x) = \prod_{s\in(0,x]}(1 - d\widehat{H}(s)) = \prod_{X_i\leq x}(1 - d\widehat{H}(X_i)),$$

por la *ecuación de Duhamel*, que es un resultado de la teoría de la integración-producto, podemos expresar la diferencia entre dos integrales-producto de la siguiente manera(Gill 1994):

$$\begin{aligned}
&\prod_{s\in(0,x]}(1 - d\widehat{H}(s)) - \prod_{s\in(0,x]}(1 - dH^*(s)) \\
&\qquad = -\int_0^x \prod_{s\in(0,u)}(1 - d\widehat{H}(s))\big(d\widehat{H}(u) - dH^*(u)\big)\prod_{s\in(u,x]}(1 - dH^*(s))
\end{aligned} \tag{1.11}$$

al tratarse de un producto de cantidades finitas, note que

$$\prod_{s\in(0,x]}(1 - dH^*(s)) = \prod_{s\in(0,u]}(1 - dH^*(s))\prod_{s\in(u,x]}(1 - dH^*(s))$$

y, por tanto que

$$\frac{\prod_{s\in(u,x]}(1 - dH^*(s))}{\prod_{s\in(0,x]}(1 - dH^*(s))} = \frac{1}{\prod_{s\in(0,u]}(1 - dH^*(s))}.$$

Entonces, tomando la diferencia del lado izquierdo en la ecuación de Duhamel (1.11), multiplicando ambos lados por $\frac{1}{\prod_{s\in(0,x]}(1-dH^*(s))}$, resulta

$$\frac{\prod_{s\in(0,x]}(1-d\widehat{H}(s)) - \prod_{s\in(0,x]}(1-dH^*(s))}{\prod_{s\in(0,x]}(1-dH^*(s))} =$$

$$\frac{\widehat{K}(x)-S^*(x)}{S^*(x)} = -\int_0^x \frac{\prod_{s\in(0,u)}(1-d\widehat{H}(s))}{\prod_{s\in(0,u]}(1-dH^*(s))}\big(d\widehat{H}(u)-dH^*(u)\big)$$

$$= -\int_0^x \frac{\widehat{K}(u-)}{S^*(u)}\big(d\widehat{H}(u)-dH^*(u)\big),$$

que es la representación en martingala del estimador Kaplan-Meier.

El lado derecho es una integral estocástica respecto de la martingala $(d\widehat{H}(u) - dH^*(u))$, lo que la hace una martingala con media 0 y, en consecuencia, $E(\frac{\widehat{K}(x)-S^*(x)}{S^*(x)}) = 0$ y $E(\frac{\widehat{K}(x)}{S^*(x)}) = 1$. Asintóticamente, $\frac{\widehat{K}(x-)}{S^*(x)} \approx 1$ y $S^*(x)$ es igual a $S(x)$, así que

$$\frac{\widehat{K}(x)}{S(x)} - 1 \approx -\int_0^x \big(d\widehat{H}(u)-dH(u)\big)$$

$$\widehat{K}(x) - S(x) \approx -S(x)\big(d\widehat{H}(x)-dH(x)\big). \tag{1.12}$$

De esta expresión, elevando al cuadrado ambos lados y tomando la esperanza se obtiene la varianza del estimador Kaplan-Meier

$$Var(\widehat{K}(x)) \approx S^2(x)Var(\widehat{H}(x)),$$

como no conocemos $S(x)$ podemos estimar la varianza con

$$\widehat{Var(\widehat{K}(x))} = \widehat{S}^2(x)\widehat{Var(\widehat{H}(x))}.$$

Por último, el término $\sqrt{n}\big(d\widehat{H}(x) - dH(x)\big)$ converge a la distribución normal a medida que $n \to \infty$ y usando (1.11) podemos afirmar que el término $\sqrt{n}\big(\widehat{K}(x)-S(x)\big)$ se distribuye normal con media 0. Se concluye que el estimador Kaplan-Meier es consistente y se distribuye asintóticamente normal.

1.3.5. Modelos de regresión

El caso simple del análisis de supervivencia supone que se estudia una población homogénea, o sea que X no está influenciada por otras variables. Un caso más probable, sin embargo, es aquel en el que X está asociada con $\mathbf{Z}^T = (Z_1, \ldots, Z_p)$ un vector de variables explicativas o covariables, que pueden ser variables cuantitativas y cualitativas.

Un primer tipo de modelo es el análogo a la regresión lineal en el que se utiliza el logaritmo natural del tiempo de falla para llevar la variable positiva del tiempo a la recta real, $Y = \ln(X)$ con $Y = \alpha + \boldsymbol{\beta}^T \boldsymbol{Z} + W$ donde $\boldsymbol{\beta}^T$ es el vector de coeficientes de regresión y W es la variable de error.

De aquí que $S(x|\boldsymbol{Z}) = P(X > x|\boldsymbol{Z}) = P(Y > \ln(x)|\boldsymbol{Z}) = P(\alpha + W > \ln(x) - \boldsymbol{\beta}^T \boldsymbol{Z}|\boldsymbol{Z}) = P(\alpha + W > x\exp(-\boldsymbol{\beta}^T \boldsymbol{Z})|\boldsymbol{Z}) = S_0(x\exp(-\boldsymbol{\beta}^T \boldsymbol{Z}))$. S_0 es la función de supervivencia de referencia que corresponde a $\boldsymbol{Z} = \mathbf{0}$, por la relación $h(x) = f(x)/S(x)$ y la regla de la cadena, su correspondiente función de riesgo es $h(x|\boldsymbol{Z}) = h_0(x\exp(-\boldsymbol{\beta}^T \boldsymbol{Z}))\exp(-\boldsymbol{\beta}^T \boldsymbol{Z})$.

Sin embargo, el uso de este primer tipo de modelos está limitado por la suposición en la distribución del error en la regresión. Otro enfoque es modelar directamente la función de riesgo como función de covariables, ya sea a través de un modelo de *riesgo multiplicativo* o un modelo *aditivo*. El primero se puede definir por $h(x|\boldsymbol{Z}) = h_0(x)c(\boldsymbol{\beta}^T \boldsymbol{Z})$ donde $c(\boldsymbol{\beta}^T \boldsymbol{Z})$ es una función liga no negativa, popularmente $c(\boldsymbol{\beta}^T \boldsymbol{Z}) = \exp(\boldsymbol{\beta}^T \boldsymbol{Z})$. (J. P. Klein y Moeschberger 2003, sección 2.6)

El segundo caso es de interés especial para este trabajo, propuesto por Aalen, de quien tomamos la siguiente descripción introductoria (O. Aalen, Borgan y Gjessing 2008, sección 4.2.1). El modelo aditivo se puede definir de forma general por

$$h(x|\boldsymbol{Z}) = \beta_0(x) + \boldsymbol{\beta}^T(x)\boldsymbol{Z}$$

donde $\beta_0(x) = h_0(x)$ y las $\beta_k(x)$ del vector $\boldsymbol{\beta}^T(x)$ pueden ser funciones arbitrarias que indican el aumento de riesgo por incremento en una unidad del tiempo.

Las covariables $\boldsymbol{Z}$ son predecibles, lo que quiere decir que se conocen desde el tiempo $x = 0$ y no varían a lo largo del tiempo de estudio. En el caso más elaborado, en el que $\boldsymbol{Z}(x)$, los valores de las covariables se conocen en $x-$.

Este modelo tiene una sencilla compatibilidad con los procesos de conteo martingala por su linealidad. Sean las funciones de regresión acumuladas $B_k(x) = \int_0^x \beta_k(u)du$

por lo que

$$\phi_i(x) = R_i(x)\Big(\beta_0(x) + \beta_1(x)Z_{i1} + \beta_2(x)Z_{i2} + \cdots + \beta_p(x)Z_{ip}\Big).$$

La regresión se hace para $dN_i(x)$ cumpliendo la conocida igualdad $dN_i(x) = \phi_i(x)dx + dM_i(x)$ con $\phi_i(x)dx = R_i(x)dB_0(x) + R_i(x)\Big(dB_1(x)Z_{i1} + dB_2(x)Z_{i2} + \cdots + dB_p(x)Z_{ip}\Big)$. Por la linealidad de estas ecuaciones es práctico expresarlas en forma de vectores y matrices: llamemos a los vectores $\boldsymbol{N}(x) = (N_1(x), \ldots, N_n(x))^T$, $\boldsymbol{M}(x) = (M_1(x), \ldots, M_n(x))^T$, $\boldsymbol{B}(x) = (B_0(x), \ldots, B_p(x))^T$ y la matriz de diseño, $\boldsymbol{X}(x)$ de dimensión $(n\cdot(p+1))$ que tiene como renglones los vectores $(R_i(x), R_i(x)Z_{i1}, \ldots, R_i(x)Z_{ip})$. Entonces

$$d\boldsymbol{N}(x) = \boldsymbol{X}(x)d\boldsymbol{B}(x) + d\boldsymbol{M}(x).$$

Si la matriz $\boldsymbol{X}(x)$ es de rango completo, es decir si sus columnas son todas linealmente independientes entre sí, podemos utilizar mínimos cuadrados ordinarios para obtener los estimadores, que resultan de

$$d\widehat{\boldsymbol{B}}(x) = (\boldsymbol{X}(x)^T\boldsymbol{X}(x))^{-1}\boldsymbol{X}(x)^T d\boldsymbol{N}(x).$$

Para un fácil manejo denotemos $\boldsymbol{X}^-(x) = (\boldsymbol{X}(x)^T\boldsymbol{X}(x))^{-1}\boldsymbol{X}(x)^T$ y a $J(x)$ como función indicadora de que $\boldsymbol{X}(x)$ es de rango completo, de modo que

$$\widehat{\boldsymbol{B}}(x) = \int_0^x J(u)\boldsymbol{X}^-(u)d\boldsymbol{N}(u) = \sum_{X_j \le x} J(X_j)\boldsymbol{X}^-(X_j)d\boldsymbol{N}(X_j).$$

Si luego denotamos

$$\boldsymbol{B}^*(x) = \int_0^x J(u)d\boldsymbol{B}(u)$$

con $\boldsymbol{B}^*(x)$ casi igual a $\boldsymbol{B}(x)$ si la probabilidad de que $\boldsymbol{X}(x)$ tenga rango completo para toda $u \in [0, x]$ es cercana a 1. Podemos utilizar los métodos para martingalas al ver que

$$\begin{aligned}
\widehat{\boldsymbol{B}}(x) - \boldsymbol{B}^*(x) &= \int_0^x J(u)\boldsymbol{X}^-(u)d\boldsymbol{N}(u) - \int_0^x J(u)\boldsymbol{I}d\boldsymbol{B}(u) \\
&= \int_0^x J(u)\boldsymbol{X}^-(u)d\boldsymbol{N}(u) - \int_0^x J(u)(\boldsymbol{X}^-(u)\boldsymbol{X}(u))d\boldsymbol{B}(u) \\
&= \int_0^x J(u)\boldsymbol{X}^-(u)d\boldsymbol{M}(u)
\end{aligned}$$

(con $\boldsymbol{I}$ como matriz de identidad) es un vector de dimensión $(p+1)$ de integrales estocásticas respecto a las martingalas $\boldsymbol{M}$, es decir el *k-ésimo* elemento de dicho vector

es $\sum_{i=1}^{n}\int_0^x \mathcal{H}_{ki}(u)dM_i$, donde $\mathcal{H}_{ki}$, el elemento $\mathbf{x}_{ki}$ de la matriz $\boldsymbol{X}^-(u)$, es un proceso estocástico predecible. Por ello mismo, es un vector de martingalas de media 0 en sí mismo, así que $E(\widehat{\boldsymbol{B}}(x) - \boldsymbol{B}^*(x)) = \mathbf{0}$.

El proceso correspondiente de variación predecible se denota por

$$\langle \widehat{\boldsymbol{B}}(x) - \boldsymbol{B}^*(x)\rangle = \int_0^x J(u)\boldsymbol{X}^-(u)diag\{\boldsymbol{\phi}(u)du\}\boldsymbol{X}^-(u)^T,$$

donde $\boldsymbol{\phi}(u) = (\phi_1(u), \ldots, \phi_n(u))$ y $diag\{\cdot\}$ hace referencia a una matriz diagonal cuyos elementos fuera de la diagonal principal son todos 0.

En cuanto a resultados asintóticos, como puede imaginarse, $\sqrt{n}(\widehat{\boldsymbol{B}}(x) - \boldsymbol{B}^*(x))$ converge en distribución a un vector de martingalas gaussianas multivariantes de media 0 y como asintóticamente $\boldsymbol{B}^*(x) = \boldsymbol{B}(x)$, $\sqrt{n}(\widehat{\boldsymbol{B}}(x) - \boldsymbol{B}(x))$ converge a la misma martingala gaussiana multivariante.

Estimación de la curva de supervivencia

El modelo aditivo puede servir para estimar indirectamente la función de supervivencia a través de la estimación de la función acumulada de riesgo condicionada a un vector de covariables. Por ejemplo, dado el vector fijo de covariables $\boldsymbol{Z}_0$ cuyo primer elemento es 1 y $\boldsymbol{B}$(x) como antes, la función acumulada de riesgo condicionada al vector de covariables es

$$H(x|\boldsymbol{Z}_0) = \int_0^x h(u|\boldsymbol{Z}_0)du = \boldsymbol{Z}_0^T\boldsymbol{B}(x),$$

y el correspondiente estimador

$$\widehat{H}(x|\boldsymbol{Z}_0) = \boldsymbol{Z}_0^T\widehat{\boldsymbol{B}}(x).$$

A través de este estimador, la función de supervivencia condicionada al vector $\boldsymbol{Z}_0$ se puede estimar con la fórmula de tipo Kaplan-Meier

$$\widehat{K}^{z_0}(x) = \widehat{S}(x|\boldsymbol{Z}_0) = \prod_{X_i \le x}(1 - d\widehat{H}(X_i|\boldsymbol{Z}_0)),$$

o equivalentemente

$$\widehat{S}(x|\boldsymbol{Z}_0) = e^{\widehat{H}(x|\boldsymbol{Z}_0)}.$$

El estimador $\widehat{K}^{Z_0}(x) = \prod_{X_i \leq x}(1 - d\widehat{H}(X_i|\boldsymbol{Z}_0))$ también cumple propiedades asintóticas similares a las del estimador Kaplan-Meier convencional. Siendo $\boldsymbol{B}^*(x) = \int_0^x J(u)d\boldsymbol{B}(u)$ como antes, notemos que

$$\begin{aligned}\sqrt{n}(\widehat{H}(x|\boldsymbol{Z}_0) - H^*(x|\boldsymbol{Z}_0)) =& \sqrt{n}[\boldsymbol{Z}_0^T(\widehat{\boldsymbol{B}}(x) - \boldsymbol{B}^*(x))] \\ =& \int_0^x \sqrt{n}J(u)\boldsymbol{Z}_0^T\boldsymbol{X}^-(u)d\boldsymbol{M}(u)\end{aligned} \tag{1.13}$$

sigue siendo un vector de integrales estocásticas de media $\mathbf{0}$, que converge en distribución a un vector de martingalas gaussianas de media $\mathbf{0}$, cuya función de varianza se puede estimar con $\boldsymbol{Z}_0^T\widehat{\boldsymbol{\Sigma}}(x)\boldsymbol{Z}_0$, donde $\widehat{\boldsymbol{\Sigma}}(x) = \sum_{X_j \leq x} J(X_j)\boldsymbol{X}^-(X_j)diag\{d\boldsymbol{N}(X_j)\}\boldsymbol{X}^-(X_j)^T$ o, según el modelo, $\widehat{\boldsymbol{\Sigma}}_{mod}(x) = \sum_{X_j \leq x} J(X_j)\boldsymbol{X}^-(X_j)diag\{\boldsymbol{X}(X_j)d\widehat{\boldsymbol{B}}(X_j)\}\boldsymbol{X}^-(X_j)^T$.

Directamente de esto podemos, como en el caso del estimador Kaplan-Meier ordinario, formar una representación martingala del nuevo estimador de la curva de supervivencia

$$\begin{aligned}\frac{\widehat{K}^{Z_0}(x) - S^*(x|\boldsymbol{Z}_0)}{S^*(x|\boldsymbol{Z}_0)} &= -\int_0^x \frac{\prod_{s\in(0,u)}(1 - d\widehat{H}(s|\boldsymbol{Z}_0))}{\prod_{s\in(0,u]}(1 - dH^*(s|\boldsymbol{Z}_0))}\big(d\widehat{H}(u|\boldsymbol{Z}_0) - dH^*(u|\boldsymbol{Z}_0)\big) \\ &= -\int_0^x \frac{\widehat{K}^{Z_0}(u-)}{S^*(u|\boldsymbol{Z}_0)}\big(d\widehat{H}(u|\boldsymbol{Z}_0)) - dH^*(u|\boldsymbol{Z}_0))\big),\end{aligned} \tag{1.14}$$

cumpliéndose como en el caso ordinario la consistencia y la normalidad asintótica.

Estimación de $S(x)$ bajo censura dependiente de covariables

Cuando ocurre que T_i y C_i no son independientes y su dependencia se puede explicar a través de un vector de covariables $\boldsymbol{Z_i}$, la estimación de la función de supervivencia no se puede dar a través del estimador Kaplan-Meier habitual, porque podría resultar, en algunos casos, significativamente sesgado.

La estimación de la función de supervivencia a través del modelo aditivo de Aalen puede presentar algunas complicaciones en la práctica debido a que varias de las matrices del tipo $(\boldsymbol{X}(x)^T\boldsymbol{X}(x))$ no tienen inversa; esto es, $\widehat{\boldsymbol{B}}(x)$ puede ser ineficiente y sesgada si varias de las matrices $\boldsymbol{X}(x)$ son de rango incompleto.

Para salvar este inconveniente, en el caso de censura dependiente de covariables, Satten, Datta y J. Robins 2001 proponen un modelo en el que estiman $S(x)$ a través

de la fórmula

$$S_C^{Z_i}(x) = \prod_{u \leq x} [1 - dH_C(u|\boldsymbol{Z}_i)]$$

en un esquema que llama de "reponderación", donde

$$\bar{N}(x) = \sum_{i=1}^{n} \frac{I(X_i \leq x, \delta_i = 1)}{S_C^{Z_i}(X_i-)}$$

y

$$\bar{Y}(x) = \sum_{i=1}^{n} \frac{I(X_i \geq x)}{S_C^{Z_i}(x-)}$$

son estimadores insesgados de $N^*(x) = \sum_{i=1}^{n} I(T_i \leq x)$ y $Y^*(x) = \sum_{i=1}^{n} I(T_i \geq x)$.

Con estos estimadores ponderados, siempre que se conozca $S_C^{Z_i}(x)$, se podría construir el estimador consistente

$$\bar{S}(x) = \prod_{X_j \leq x} [1 - \frac{d\bar{N}(X_j)}{\bar{Y}(X_j)}].$$

Cuando no se conoce $S_C^{Z_i}(x-)$, Satten, Datta y J. Robins 2001 proponen estimarla de la siguiente manera. El objetivo es estimar $H_C(x|\boldsymbol{Z}_i) = \int_0^x \boldsymbol{Z}_i^T \, d\boldsymbol{B}(u)$. $\boldsymbol{B}(x)$ se estima por

$$\widehat{\boldsymbol{B}}(x) = \sum_{i=1}^{n} I(X_i \leq x)(1 - \delta_i)\boldsymbol{A}^{-1}(x)\boldsymbol{Z}_i,$$

donde

$$\boldsymbol{A}(x) = \sum_{i=1}^{n} I(X_i \geq x)\boldsymbol{Z}_i\boldsymbol{Z}_i^T.$$

En los casos en que $\boldsymbol{A}(x)$ no sea invertible, el autor indica que el cálculo se puede hacer con una matriz inversa generalizada, él propone en su artículo el uso de la *inversa espectral* $P \cdot Diag(E^+) \cdot P^T$, donde P es la matriz cuyas columnas son los vectores propios de $\boldsymbol{A}(x)$, P^T es su transpuesta y E^+ es una matriz diagonal cuyas entradas $E_{i,i}$ son iguales a los recíprocos de los valores propios de $\boldsymbol{A}(x)$ cuando éstos son diferentes de 0, y 0 para los valores propios iguales a 0.

Capítulo 2

Métodos y obtención de estimadores

2.1. Métodos

2.1.1. Método de ponderación de probabilidad inversa

La ponderación de probabilidad inversa es una técnica estadística aplicada primeramente por Horvitz y Thompson 1952 en su trabajo *"A generalization of sampling without replacement from a finite universe"*, para estimar la media de una población a través de una muestra estratificada y las probabilidades de que un elemento pertenezca a un determinado estrato, es conocido como el estimador Horvitz-Thompson.

En general, se puede decir que sirve para calcular estadísticos estandarizados donde hay restricciones a la recolección de datos por diferentes motivos y algunos de ellos no pueden ser registrados. La solución propone un diseño estratificado de la muestra de tal forma que a cada estrato le corresponda una probabilidad para usarla como ponderación.

Si se conoce la probabilidad a partir de la cual la población de muestreo es extraída de la población objetivo, se usa la inversa de esta probabilidad para ponderar las observaciones. El enfoque ha sido generalizado y es utilizado, además del cálculo de la media, en funciones de densidad o de verosimilitud.

Cuando hay muchos datos faltantes, el método de ponderación de probabilidad inversa puede ayudar a reducir el peso de elementos sobre representados o, a la inversa,

aumentar el de aquellos individuos cuyos datos sean los de mayor dificultad de acceso.

Para entender este método de forma sencilla se puede utilizar un ejemplo del estimador Horvitz-Thompson (Zhu 2012). Supongamos que tenemos los siguientes datos:

Grupo	A			B			C		
Respuesta	3	3	3	6	6	6	9	9	9

El promedio de la respuesta es 6. Pero si sólo observáramos los siguientes datos

Grupo	A			B			C		
Respuesta	3	?	?	6	6	6	?	9	9

El promedio sería 6.5, que es un resultado sesgado. Ahora, la probabilidad de respuesta es 1/3 en el grupo A, 1 en el grupo B y 2/3 en el grupo C. Así que podemos calcular un promedio ponderado, en el que cada observación es ponderada por 1/Probabilidad de respuesta, dividiendo por la suma de las probabilidades inversas de cada observación:

$$\frac{3 * \frac{3}{1} + (6+6+6) * 1 + (9+9) * \frac{3}{2}}{\frac{3}{1} + 1 + 1 + 1 + \frac{3}{2} + \frac{3}{2}}$$

Obteniendo como resultado nuevamente 6, insesgado. Más generalmente, este método puede devolver estimaciones sesgadas de los parámetros, pero serán las menos sesgadas.

2.1.2. Método delta

Las funciones de variables aleatorias son también variables aleatorias y su distribución de probabilidad depende de la distribución de las variables originales. El método delta puede considerarse junto con el del teorema central del límite como un resultado que describe las propiedades asintóticas de estimadores y de funciones de estimadores.

En suma, el método delta afirma que si se tiene una sucesión de variables aleatorias $\{X_n\}$ con media μ y varianza σ^2, y se satisface que $\frac{\sqrt{n}(X_n-\mu)}{\sigma} \xrightarrow{d} N(0,1)$; esto es, la estandarización de esta sucesión converge en distribución a la normal estándar, entonces para cualquier función $g(\cdot)$, tal que $g'(\mu)$ exista y sea diferente de 0, se cumple que

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{\sigma|g'(\mu)|} \xrightarrow{d} N(0,1)$$

o

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2(g'(\mu))^2).$$

Series de Taylor

Para las aplicaciones estadísticas bastan algunos términos de la serie de Tylor, de hecho en la mayoría es suficiente con la serie de Tylor de primer orden, la que se expande sólo hasta la primera derivada con $k = 1$.(Casella y Berger 2002, p. 241)

En series de Taylor estocásticas, dada una secuencia aleatoria X_n tal que $X_n = a + O_p(r_n)$ y una función g que tenga derivadas continuas cerca de a hasta el orden $s + 1$, entonces se cumple que

$$g(X_n) = \sum_{i=0}^{s} \frac{g^i(a)}{i!}(X_n - a)^i + o_p(r_n)$$

donde $g(X_n)$ es también una secuencia aleatoria.(Kowalski y Tu 2008, p.52-53)

Dado un conjunto de variables aleatorias $\mathbf{T} = (T_1, T_2, \ldots, T_m)$ cuyas medias sean $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_m$, tomando $\boldsymbol{\theta}$ como punto para la expansión, tenemos que

$$g_i'(\boldsymbol{\theta}) = \frac{\partial}{\partial t_i} g(\mathbf{t})|_{t_1=\theta_1,\ldots,t_m=\theta_m}$$

así que la aproximación de $g(\mathbf{t})$ por serie de Tylor de primer orden queda de la siguiente forma

$$g(\mathbf{t}) \approx g(\boldsymbol{\theta}) + \sum_{i=1}^{m} g_i'(\boldsymbol{\theta})(t_i - \theta_i).$$

Estimación de la varianza

Si calculáramos la esperanza de ambos lados de la expresión anterior obtendríamos que

$$\begin{aligned} E[g(\mathbf{t})] &\approx E[g(\boldsymbol{\theta}) + \sum_{i=1}^{m} g_i'(\boldsymbol{\theta})(t_i - \theta_i)] \\ &= g(\boldsymbol{\theta}) + \sum_{i=1}^{m} g_i'(\boldsymbol{\theta}) E[(t_i - \theta_i)] \\ &= g(\boldsymbol{\theta}). \end{aligned}$$

De ahí que la varianza se aproxime con

$$\begin{aligned} Var[g(\mathbf{t})] &\approx E[(g(\mathbf{t}) - g(\boldsymbol{\theta}))^2] \\ &= E[(g(\mathbf{t}))^2] - g^2(\boldsymbol{\theta}) \end{aligned}$$

y como $g(\mathbf{t}) \approx g(\boldsymbol{\theta}) + \sum_{i=1}^{m} g_i'(\boldsymbol{\theta})(t_i - \theta_i)$

$$\begin{aligned} Var[g(\mathbf{t})] &\approx E[(g(\boldsymbol{\theta}) + \sum_{i=1}^{m} g_i'(\boldsymbol{\theta})(t_i - \theta_i))^2] - g(\boldsymbol{\theta})^2 \\ &= g(\boldsymbol{\theta})^2 + 2g(\boldsymbol{\theta}) \sum_{i=1}^{m} g_i'(\boldsymbol{\theta}) E[(t_i - \theta_i)] + E[(\sum_{i=1}^{m} g_i'(\boldsymbol{\theta})(t_i - \theta_i))^2] - g(\boldsymbol{\theta})^2 \\ &= E[(\sum_{i=1}^{m} g_i'(\boldsymbol{\theta})(t_i - \theta_i))^2] \\ &= \sum_{i=1}^{m} g_i'(\boldsymbol{\theta})^2 E[(t_i - \theta_i)^2] + 2 \sum_{1 \le i < j \le m} g_i'(\boldsymbol{\theta}) g_j'(\boldsymbol{\theta}) E[(t_i - \theta_i)(t_j - \theta_j)] \\ &= \sum_{i=1}^{m} g_i'(\boldsymbol{\theta})^2 Var(T_i) + 2 \sum_{1 \le i < j \le m} g_i'(\boldsymbol{\theta}) g_j'(\boldsymbol{\theta}) Cov(T_i, T_j) \end{aligned}$$

2.1.3. *Estadísticos-U*

Pensemos en un funcional θ definido en que es el conjunto de todas las funciones de distribución F en $\mathbb{R}$. Nuestro propósito es estimar $\theta = \theta(F)$, $F \in \Theta$, a partir de una muestra aleatoria $T_1, \ldots, T_n$ con función de distribución F. (Lee 1990)

Una función $\varphi(t_1, \ldots, t_m)$, con $n \ge m$, será el estimador insesgado del funcional θ si $\theta(F) = E(\varphi(T_1, \ldots, T_m)) = \int \cdots \int \varphi(t_1, \ldots, t_m) dF(t_1) \cdots dF(t_m)$. $\varphi(t_1, \ldots, t_m)$ es comúnmente denominado *kernel* y nos interesa uno simétrico, es decir que su valor no varíe ante las permutaciones de sus elementos.

El estimador

$$\widehat{\theta} = U_n = \binom{n}{m}^{-1} \sum_{(n,m)} \varphi(T_{i1}, \ldots, T_{im})$$

es el correspondiente estadístico-U, donde $\sum_{(n,m)} = \sum_{1\leq i_1<\cdots<i_m\leq n}$. Dos ejemplos sencillos para entender lo que estamos tratando son la media muestral, donde $m = 1$

$$\overline{T} = U_n = \frac{1}{n}\sum_{i=1}^{n} T_i$$

y la varianza muestral, donde $m = 2$

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(T_i - \overline{T})^2 = U_n = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}\frac{1}{2}(T_i - Tj)^2.$$

Los estadísticos-U son claramente insesgados, pero no sólo eso, sino que son estimadores insesgados de varianza mínima (Serfling 1980, p.176).

Proyección de estadísticos-U

Para los estadísticos-U con $m > 1$, los elementos sumados no son todos independientes entre sí, esta condición impide la aplicación de los resultados conocidos para las sumas de variables *i.i.d*, como la Ley de los Grandes Números o el Teorema del Límite Central. La proyección para estadísticos-U permite la aproximación de estos estadísticos por medio de otra suma de variables *i.i.d.* a la que sí son aplicables los resultados propios de estas sumas, como la teoría asintótica.

La proyección del estadístico-U se define como

$$\widehat{U}_n = \sum_{i=1}^{n} E(U_n|T_i) - (n-1)\theta$$

$\widehat{U}_n$ es la proyección de U_n sobre cada T_i. Y llamemos *proyección centrada* a

$$\widehat{U}_n - \theta = \sum_{i=1}^{n} E(U_n|T_i) - n\theta = \sum_{i=1}^{n}[E(U_n|T_i) - \theta].$$

Sea

$$E[\varphi(T_1,\ldots,T_m)|T_i] = \begin{cases}\varphi_1(T_i) & \text{if } i \in \{1,\ldots,m\}\\ \theta & \text{if } i \notin \{1,\ldots,m\}\end{cases},$$

note que (Kowalski y Tu 2008, p. 154)

$$\begin{aligned} E(U_n|T_i) &= \binom{n}{m}^{-1}\left[\sum_{(n,m)}\left(\sum_{i\in\{1,\ldots,m\}} E(\varphi|T_i) + \sum_{i\notin\{1,\ldots,m\}} E(\varphi|T_i)\right)\right] \\ &= \binom{n}{m}^{-1}\left[\binom{n-1}{m-1}\varphi_1(T_i) + \binom{n-1}{m}\theta\right] \\ &= \frac{m}{n}\varphi_1(T_i) + \frac{n-m}{n}\theta. \end{aligned}$$

Con $\tilde{\varphi}_1(T_i) = \varphi_1(T_i) - \theta$, la proyección centrada se vuelve

$$\begin{aligned} \widehat{U}_n - \theta =& \sum_{i=1}^{n}[E(U_n|T_i) - \theta] = \sum_{i=1}^{n}[\frac{m}{n}\varphi_1(T_i) + \frac{n-m}{n}\theta - \theta] \\ =& \sum_{i=1}^{n}[\frac{m}{n}\varphi_1(T_i) - \frac{m}{n}\theta] \\ =& \frac{m}{n}\sum_{i=1}^{n}\tilde{\varphi}_1(T_i). \end{aligned}$$

Antes de presentar los resultados asintóticos, presentaremos un ejemplo que será útil para este trabajo por su representación multivariante, el de la proyección del estadístico-U para la covarianza. Primero recordemos que el estimador clásico insesgado para la covarianza es

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Para el par de variables (X,Y), el estadístico-U es

$$U_n = \sum_{(n,2)}\varphi(\boldsymbol{Z_i}, \boldsymbol{Z_j}) = \frac{1}{2}(X_i - X_j)(Y_i - Y_j),$$

donde $\boldsymbol{Z_i} = (X_i, Y_i)^T$. Por lo que

$$\begin{aligned}
\varphi_1(\boldsymbol{Z_i}) &= E(\varphi(\boldsymbol{Z_i}, \boldsymbol{Z_j})|\boldsymbol{Z_i}) = E(\frac{1}{2}(X_i - X_j)(Y_i - Y_j)|\boldsymbol{Z_i}) \\
&= \frac{1}{2}[x_i y_i - x_i\mu_Y - \mu_X y_i + E(XY) + \mu_X\mu_Y - \mu_X\mu_Y] \\
&= \frac{1}{2}[(x_i - \mu_X)(y_i - \mu_Y) + (E(XY) - \mu_X\mu_Y)] \\
&= \frac{1}{2}[(x_i - \mu_X)(y_i - \mu_Y) + E(X_jY_j - X_j\mu_Y - \mu_X Y_j + \mu_X\mu_Y)] \\
&= \frac{1}{2}[(x_i - \mu_X)(y_i - \mu_Y) + E((X_j - \mu_X)(Y_j - \mu_Y))] \\
&= \frac{1}{2}[(x_i - \mu_X)(y_i - \mu_Y) + \sigma_{XY}] \\
&= \frac{1}{2}[(x_i - \mu_X)(y_i - \mu_Y) + \theta]
\end{aligned}$$

Obteniendo como proyección centrada para este estadístico-U

$$\widehat{U}_n - \theta = \frac{2}{n}\sum_{i=1}^{n}\frac{1}{2}[(x_i - \mu_X)(y_i - \mu_Y) - \sigma_{XY}].$$

Resultados asintóticos

Por los resultados anteriores es fácil corroborar que $\widehat{U}_n$ es un estimador insesgado de θ, ya que $E[\tilde{\varphi}_1(T_i)] = E[E(\varphi(T_1, \ldots, T_m)|T_i) - \theta] = 0$. Por la ley de los grandes números y el teorema del límite central, $\widehat{U}_n \xrightarrow{p} \theta$ y $\sqrt{n}(\widehat{U}_n - \theta)$ se distribuye asintóticamente normal con media 0.

Un resultado de la teoría asintótica para estadísticos-U surgidos de una muestra (puede ver Kowalski y Tu 2008, sección 3.2.2), dice que

$$\sqrt{n}(U_n - \theta) - \sqrt{n}(\widehat{U}_n - \theta) = o_p(1),$$

que implica que $U_n - \theta$ y $\widehat{U}_n - \theta$ siguen la misma distribución asintótica, es de esta forma que se afirma que el estadístico U_n se distribuye asintóticamente normal.

El estimador definido por

$$V_n = \frac{1}{n^m}\sum_{i1=1}^{n}\cdots\sum_{im=1}^{n}\varphi(T_{i1}, \ldots, T_{im}),$$

notoriamente relacionado con el estadístico-U de m argumentos, es conocido como estadístico de *Von Mises* y es otro estimador insesgado de θ.

Esta relación tiene una importante implicación y es que, como se muestra en Serfling 1980, sección 5.7.3

$$\sqrt{n}(U_n - V_n) \xrightarrow{p} 0,$$

que a la vez implica que $\sqrt{n}(U_n - \theta)$ y $\sqrt{n}(V_n - \theta)$ siguen la misma distribución, V_n también se distribuye asintóticamente normal.

2.2. Estimación del índice de Gini con censura independiente

Partimos de la suposición de que T y C son independientes. Cada individuo tiene un tiempo T_i hasta la ocurrencia del evento de interés, conseguir un trabajo, y un tiempo de censura C_i, en caso de que durante el estudio no sea haya observado el evento de interés o haya sucedido un riesgo competitivo.

La censura por la derecha implica que para algunos individuos el tiempo T_i no es observable, lo que observamos es $X_i = min(T_i, C_i)$, así que necesitamos un estimador que sea lo más **insesgado** posible.

Sea $P(\delta = 1) = P(C > T)$ y llamemos $S_{_C}(x) = P(C > x)$. Dada la independencia entre T y C, y que δ se distribuye como una binomial con probabilidad de éxito $p = P(\delta = 1) = P(C > T) = S_{_C}(T)$, tenemos que

$$E(\delta|T) = P(\delta = 1|T) = P(C > T|T) = P(C > T) = S_{_C}(T) \tag{2.1}$$

Proposición 2. *Si la función $g(\cdot)$ está bien definida*[1]*, entonces*

$$E\left[\frac{g(X)\delta}{S_{_C}(X)}\right] = E[g(T)]. \tag{2.2}$$

Demostración:

[1]Se dice de una función que está bien definida si no hay ambigüedad en su definición de tal forma que entradas del mismo valor con distintas formas no llevan a resultados distintos. El ejemplo común de una expresión mal definida es $f(a/b) = a + b$, cuyo resultado cambia aun cuando el valor de a/b se mantenga constante.

$$
\begin{aligned}
E\left[\frac{g(X)\delta}{S_C(X)}\right] &= \int_0^\infty \sum_\delta \frac{g(X)\delta}{S_C(X)} f(X,\delta)dx \\
&= \int_0^\infty \frac{g(t)\cdot 1}{S_C(t)} f(t,1)dt + \int_0^\infty \frac{g(c)\cdot 0}{S_C(c)} f(c,0)dc \\
&= \int_0^\infty \frac{g(t)\cdot 1}{S_C(t)} f_T(t) S_C(t)dt + \int_0^\infty \frac{g(c)\cdot 0}{S_C(c)} f_C(c) S_T(c)dc \\
&= \int_0^\infty g(t) f_T(t)dt = E[g(T)]
\end{aligned}
$$

La igualdad anterior deja ver que a pesar de no poder observar todos los tiempos T_i, podemos calcular $E(g(T))$ a través de $E\left(\frac{g(X)\delta}{S_C(X)}\right)$. Esta ventaja se le debe al método de ponderación de probabilidad inversa que utiliza las ponderaciones $\frac{\delta}{S_C(x)}$ para aproximar al valor real.

Para construir el estimador del coeficiente de Gini, correspondiente a la censura independiente, usaremos el estimador Kaplan-Meier $\widehat{K}_C(u) = \prod_{x_i \le u}\left[1 - \frac{dN^C(x_i)}{Y(x_i)}\right]$ para aproximar $S_C(u)$. Una vez calculados todos los estimadores $\{\widehat{K}_C(X_i)\}_{i=1}^n$ para nuestra muestra aleatoria X_i, el estimador del coeficiente de Gini se calcula con la siguiente expresión:

$$
\widehat{G} = \frac{\widehat{\xi}}{\widehat{\mu}} - 1 \tag{2.3}
$$

donde $\widehat{\xi} = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i 2F_{nc}(X_i)X_i}{\widehat{K}_C(X_i)}$, $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i X_i}{\widehat{K}_C(X_i)}$ y $F_{nc}(x) = \frac{1}{n}\sum_{j=1}^n \frac{\delta_j I(X_j \le x)}{\widehat{K}_C(X_j)}$.

Del estimador (2.3) se espera **insesgamiento asintótico**. El insesgamiento asintótico es la propiedad del estimador $\widehat{Q}$ tal que $\lim_{n\to\infty} E(Q_n) = \varrho(\theta)$, para todo $\theta \in \Theta$, lo que quiere decir que el sesgo se reduce y el valor esperado del estimador se acerca al valor real del parámetro a medida que la muestra aleatoria es más grande y tiende a infinito. También esperamos que el estimador sea consistente.

El estimador $\widehat{G}$ cumple con estas propiedades asintóticas deseables como se mostrará en un momento. Hay tres condiciones que sostienen las propiedades asintóticas

deseables del estimador.

Suposición 1 La muestra $\{X_i, \delta_i\}_{i=1}^n$ está compuesta de variables independientes e idénticamente distribuidas (v.i.i.d.) y las variables T y C provienen de distribuciones que son continuas y positivas.

Suposición 2 Las variables T y C son independientes, los tiempos de censura son independientes del tiempo del evento principal.

Suposición 3 $Var[\sqrt{n}(\widehat{G} - G)] = \Omega$ existe.

El siguiente teorema muestra que el estimador $\widehat{G}$ es consistente y asintóticamente normal.

Teorema 1. *1. Cuando se cumplen las suposiciones 1 y 2, a medida que $n \to \infty$, sucede que $\widehat{G} = G + o_p(1)$. 2. Cuando se cumplen las 3 suposiciones, a medida que $n \to \infty$, se tiene que*

$$\sqrt{n}(\widehat{G} - G) \xrightarrow{d} N(0, \Omega)$$

donde

$$\begin{aligned}
\Omega =& \frac{1}{\mu^2}\Bigg\{ var[2V_\xi(T_i) - (G+1)T_i] + 4\Bigg[E\left(\int_0^\infty \frac{F(V_\xi,u)^2}{S_C(u)^2} h_C(u)R_i(u)du\right) \\
&+ E\left(\int_0^\infty \frac{V_\xi(T_i)^2}{S_C(u)^2} h_C(u)R_i(u)du\right) - 2E\left(\int_0^\infty \frac{F(V_\xi,u)V_\xi(T_i)}{S_C(u)^2} h_C(u)R_i(u)du\right)\Bigg] \\
&+ (G+1)^2\Bigg[E\left(\int_0^\infty \frac{F(T,u)^2}{S_C(u)^2} h_C(u)R_i(u)du\right) \\
&+ E\left(\int_0^\infty \frac{T_i^2}{S_C(u)^2} h_C(u)R_i(u)du\right) - 2E\left(\int_0^\infty \frac{F(T,u)T_i}{S_C(u)^2} h_C(u)R_i(u)du\right)\Bigg] \\
&- 4(G+1)\Bigg[E\left(\int_0^\infty \frac{F(T,u)F(V_\xi,u)}{S_C(u)^2} h_C(u)R_i(u)du\right) \\
&+ E\left(\int_0^\infty \frac{T_iV_\xi(T_i)}{S_C(u)^2} h_C(u)R_i(u)du\right) - E\left(\int_0^\infty \frac{F(T,u)V_\xi(T_i)}{S_C(u)^2} h_C(u)R_i(u)du\right) \\
&- E\left(\int_0^\infty \frac{F(V_\xi,u)T_i}{S_C(u)^2} h_C(u)R_i(u)du\right)\Bigg]\Bigg\}
\end{aligned}$$

con $F(T,u) = E[T_i I(T_i \geq u)]/S_T(u)$, $F(V_\xi, u) = E[V_\xi(T_i) I(T_i \geq u)]/S_T(u)$ *y* $V_\xi(T) = E\Big(I(T_j \leq T_i)T_i + I(T_i \leq T_j)T_j | T_i\Big) = \int_{T_i}^{\infty} u dF(u) + T_i F(T_i)$.

Demostración:

1. Para probar la consistencia del estimador $\widehat{G}$ empecemos por recordar que $\widehat{K}_{C}(x) \xrightarrow{p} S_{C}(x)$, es decir $\widehat{K}_{C}(x) = S_{C}(x) + o_p(1)$. Por lo tanto, de manera indirecta, siendo $\tilde{F}_{nc}(x)$, $\tilde{\mu}$ y $\tilde{\xi}$ los estimadores $\widehat{F}_{nc}(x)$, $\widehat{\mu}$ y $\widehat{\xi}$ pero con $S_{C}(x)$ en lugar de $\widehat{K}_{C}(x)$, resulta que

$$\begin{aligned} \widehat{F}_{nc}(x) =& \tilde{F}_{nc}(x) + o_p(1) \\ \widehat{\mu} =& \tilde{\mu} + o_p(1) \\ \widehat{\xi} =& \tilde{\xi} + o_p(1) \end{aligned}$$

y $\widehat{G} = \tilde{G} + o_p(1)$, por ser las funciones $\tilde{(\cdot)}$ continuas en $S_{C}(x)$. Aplicando la ley de los grandes números y la ley de la esperanza total tendremos ahora

$$\begin{aligned} \tilde{F}_{nc}(x) =& F_c(x) + o_p(1) \\ \tilde{\mu} =& \mu + o_p(1) \\ \tilde{\xi} =& \xi + o_p(1) \end{aligned}$$

lo que significa que $\widehat{G} = G + o_p(1) + o_p(1) = G + o_p(1)$, demostrando la consistencia del estimador.

2. Veamos ahora la normalidad del estimador y su varianza.

Por el primer término de la serie de Taylor estocástica para $\widehat{G}(\widehat{\mu}, \widehat{\xi})$ alrededor de (μ, ξ)

$$\widehat{G} = \frac{\xi}{\mu} - 1 + \frac{\widehat{\xi} - \xi}{\mu} - \frac{\xi}{\mu^2}(\widehat{\mu} - \mu) + o_p(n^{-1/2})$$

Sea la aproximación:

$$\begin{aligned}\widehat{\mu}-\mu &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i X_i}{\widehat{K}_C(X_i)}-\mu = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i}{\widehat{K}_C(X_i)}-\mu \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i}{\widehat{K}_C(X_i)}-\mu+\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i}{S_C(X_i)}-\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i}{S_C(X_i)} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i}{S_C(X_i)}-\mu\right)+\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i\left(S_C(X_i)-\widehat{K}_C(X_i)\right)}{S_C(X_i)\widehat{K}_C(X_i)} \\ &= a_{n1}+a_{n2}\end{aligned} \tag{2.4}$$

Si $M_i^C(t) = N_i^C(t) - \int_0^t R_i(u)h_C(u)du$, es el proceso de conteo martingala con respecto a C

$$dM_i^C(x) = dN_i^C(x) - R_i(x)h_C(x)dx$$

de modo que

$$\begin{aligned}1-\int_0^\infty \frac{dM_i^C(u)}{S_C(u)}du &= 1-\int_0^\infty \frac{dN_i^C(u)}{S_C(u)}du+\int_0^\infty \frac{R_i(u)h_C(u)}{S_C(u)}du \\ &= 1-\int_0^\infty \frac{dN_i^C(u)}{S_C(u)}du+\int_0^\infty \frac{I(X_i\geq u)h_C(u)}{S_C(u)}du \\ &= 1-\int_0^\infty \frac{dN_i^C(u)}{S_C(u)}du+\int_0^{X_i} \frac{h_C(u)}{S_C(u)}du \\ &= 1-\int_0^\infty \frac{dN_i^C(u)}{S_C(u)}du+\frac{1}{S_C(X_i)}-\frac{1}{S_C(0)} \\ &= \frac{1}{S_C(X_i)}-\int_0^\infty \frac{dN_i^C(u)}{S_C(u)}du.\end{aligned}$$

Debido a que N_i^C es un proceso de conteo, continuo por la derecha, no decreciente que crece a saltos cuando se cumple que $X_i = x$ y $\delta = 0$, entonces

$$\int_0^\infty \frac{dN_i^C(u)}{S_C(u)}du = \sum_{X_j}\frac{dN_i^C(X_j)}{S_C(X_j)}$$

esto es igual a $\frac{0}{S_C(X_j)}$ si $\delta = 1$, o igual $\frac{1}{S_C(X_j)}$ si $\delta = 0$, ya que la suma solo puede tener un elemento diferente de 0, que será $dN_i^C(X_j) = I(X_i = X_j, \delta_i = 0)$. Así, podemos como en J. M. Robins y Rotnitzky 1992, (sección 3h), expresar la equivalencia

$$\frac{\delta_i}{S_C(X_i)} = 1-\int_0^\infty \frac{dM_i^C(u)}{S_C(u)}. \tag{2.5}$$

Sustituimos (2.5) en a_{n1}

$$\frac{1}{n}\sum_{i=1}^{n} T_i - \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty} \frac{T_i}{S_{_C}(u)} dM_i^{C}(u) - \mu. \tag{2.6}$$

Fijémonos en a_{n2}, este término tiene como factor la representación martingala del estimador Kaplan-Meier

$$\begin{aligned}
\frac{S_{_C}(x) - \widehat{K}_{_C}(x)}{S_{_C}(x)} &= \int_0^x \frac{\widehat{K}_{_C}(u-)}{S_{_C}(u)}\left(\frac{dN^{C}(u)}{R(u)} - h_{_C}(u)du\right) \\
&= \int_0^x \frac{\widehat{K}_{_C}(u-)}{S_{_C}(u)}\left(\frac{dM^{C}(u)}{R(u)}\right) \\
&= \int_0^{\infty} \frac{I(x \geq u)\widehat{K}_{_C}(u-)}{S_{_C}(u)}\left(\frac{dM^{C}(u)}{R(u)}\right).
\end{aligned}$$

Vimos que $N^{C}(x) = \sum_{i=1}^{n} I(X_i \leq x, \delta = 0)$, entonces $N^{T}(x) = \sum_{i=1}^{n} I(X_i \leq x, \delta = 1)$ y $n = N^{C}(x) + N^{T}(x) + R(x+)$, donde $R(x+) = \sum_{i=1}^{n} I(X_i > x)$, así como $N^{C,T}(x-) = \sum_{i=1}^{n} I(X_i < x, \delta)$. Verificamos que

$$\begin{aligned}
\widehat{K}_{_C}(u-)\widehat{K}_{T}(u-) &= \prod_{s\in(0,u)}\left(1 - \frac{dN^{C}(s)}{R(s)}\right)\left(1 - \frac{dN^{T}(s)}{R(s)}\right) \\
&= \prod_{s\in(0,u)}\left(1 - \frac{dN^{C}(s)}{R(s)} - \frac{dN^{T}(s)}{R(s)} + \frac{dN^{C}(s)dN^{T}(s)}{R(s)^2}\right)
\end{aligned}$$

(el último término de esta expresión siempre es igual a 0 debido a que $dN(u)$ solo puede ser igual a 1 o 0, a la vez que si $dN^{C}(u) = 1$ entonces $dN^{T}(u) = 0$ y

viceversa.)

$$
\begin{aligned}
&= \prod_{s\in(0,u)} \left(1 - \frac{dN^C(s) + dN^T(s)}{R(s)}\right) \\
&= \prod_{s\in(0,u)} \left(1 - \frac{N^C(s) - N^C(s-) + N^T(s) - N^T(s-)}{R(s)}\right) \\
&= \prod_{s\in(0,u)} \left(\frac{R(s) - n + R(s+) + n - R(s)}{R(s)}\right) \\
&= \prod_{s\in(0,u)} \left(\frac{R(s+)}{R(s)}\right) = \left(\frac{R(s_1+)}{n}\right)\cdots\left(\frac{R(u)}{R(u-)}\right) \\
\widehat{K}_C(u-)\widehat{K}_T(u-) &= \frac{R(u)}{n}.
\end{aligned}
$$

Así que $\frac{\widehat{K}_C(u-)}{R(u)} = \frac{1}{n\widehat{K}_T(u-)}$. Retomamos la ecuación de la representación en martingala del estimador y sustituimos

$$
\frac{S_C(X_i) - \widehat{K}_C(X_i)}{S_C(X_i)} = \int_0^\infty \frac{I(X_i \geq u)}{n\widehat{K}_T(u-)S_C(u)} dM^C(u). \tag{2.7}
$$

Sustituimos (2.7) en a_{n2} para tener

$$
\begin{aligned}
a_{n2} &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i X_i}{\widehat{K}_C(X_i)} \int_0^\infty \frac{I(X_i \geq u)}{n\widehat{K}_T(u-)S_C(u)} dM^C(u) \\
&= \int_0^\infty \left[\frac{1}{n}\sum_{i=1}^n \frac{\delta_i X_i I(X_i \geq u)}{\widehat{K}_C(X_i)\widehat{K}_T(u-)}\right] \frac{dM^C(u)}{nS_C(u)} \\
&= \int_0^\infty \frac{\widehat{F}(T,u)}{nS_C(u)} dM^C(u)
\end{aligned}
$$

(donde $\widehat{F}(T,u)$ es el estimador ponderado de $F(T,u)$. Sabemos que $\widehat{F}(T,u) = F(T,u) + o_p(1)$, como en la primera parte de la demostración, de manera indirecta por convergencia a una funciones idéntica pero sustituyendo $S_C(X_i)$ y $S_T(u)$, y aplicando la ley de los grandes números y la ley de la esperanza total. Luego, aplicando el método delta como generalización del teorema del límite central,

tenemos...)

$$\begin{aligned} &= \int_0^\infty \frac{F(T,u)}{nS_C(u)} dM^C(u) + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^n \int_0^\infty \frac{F(T,u)}{S_C(u)} dM_i^C(u) + o_p(n^{-1/2}). \end{aligned} \tag{2.8}$$

Retomamos (2.4) y sustituimos en ella (2.6) y (2.8)

$$\begin{aligned} \widehat{\mu} - \mu =& \frac{1}{n}\sum_{i=1}^n T_i - \frac{1}{n}\sum_{i=1}^n \int_0^\infty \frac{T_i}{S_C(u)} dM_i^C(u) - \mu \\ &+ \frac{1}{n}\sum_{i=1}^n \int_0^\infty \frac{F(T,u)}{S_C(u)} dM_i^C(u) + o_p(n^{-1/2}). \end{aligned} \tag{2.9}$$

Ahora fijémonos en la aproximación $\widehat{\xi} - \xi$

$$\begin{aligned} \widehat{\xi} - \xi &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i 2F_{nc}(X_i)X_i}{\widehat{K}_C(X_i)} - \xi = \frac{1}{n^2}\sum_{i=1}^n \frac{\delta_i 2X_i}{\widehat{K}_C(X_i)}\left(\sum_{j=1}^n \frac{\delta_j I(X_j \le X_i)}{\widehat{K}_C(X_j)}\right) - \xi \\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{\delta_i\delta_j 2T_i I(T_j \le T_i)}{\widehat{K}_C(X_i)\widehat{K}_C(X_j)} - \xi \end{aligned}$$

(note que $2T_iI(T_j \le T_i)$ puede sustituirse por la función simétrica respecto de (i,j), $\boldsymbol{V_{ij} = I(T_j \le T_i)T_i + I(T_i \le T_j)T_j}$)

$$\begin{aligned} =& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{\delta_i\delta_j V_{ij}}{\widehat{K}_C(X_i)\widehat{K}_C(X_j)} - \xi \\ =& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{\delta_i\delta_j V_{ij}}{\widehat{K}_C(X_i)\widehat{K}_C(X_j)} - \xi + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{\delta_i\delta_j V_{ij}}{S_C(X_i)S_C(X_j)} \\ &- \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{\delta_i\delta_j V_{ij}}{S_C(X_i)S_C(X_j)} \\ =& \left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{\delta_i\delta_j V_{ij}}{S_C(X_i)S_C(X_j)} - \xi\right] \\ &+ \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{\delta_i\delta_j V_{ij}\left(S_C(X_i)S_C(X_j) - \widehat{K}_C(X_i)\widehat{K}_C(X_j)\right)}{S_C(X_i)S_C(X_j)\widehat{K}_C(X_i)\widehat{K}_C(X_j)} \\ =& b_{n1} + b_{n2} \end{aligned} \tag{2.10}$$

La primera parte, b_{1n}, es un estadístico de Von-Mises, con $\boldsymbol{l_i} = (X_i, \delta_i)$, menos el parámetro que aproxima. Asintóticamente, podemos utilizar la teoría de los estadísticos-U y aplicar una proyección centrada a b_{n1}

$$\begin{aligned} b_{n1} &= \frac{2}{n}\sum_{i=1}^{n}\left[E\left(\frac{\delta_i\delta_j V_{ij}}{S_C(X_i)S_C(X_j)}\middle|\boldsymbol{l_i}\right) - \xi\right] + o_p(n^{-1/2}) \\ &= \frac{2}{n}\sum_{i=1}^{n}\left[\frac{\delta_i}{S_C(X_i)}E(V_{ij}|\boldsymbol{l_i}) - \xi\right] + o_p(n^{-1/2}) \\ &= \frac{2}{n}\sum_{i=1}^{n}\left[\frac{\delta_i V_\xi(T_i)}{S_C(X_i)} - \xi\right] + o_p(n^{-1/2}). \end{aligned} \tag{2.11}$$

En esta expresión también podemos sustituir (2.5)

$$b_{n1} = \frac{2}{n}\sum_{i=1}^{n}\Big(V_\xi(T_i) - \xi\Big) - \frac{2}{n}\sum_{i=1}^{n}\int_0^\infty \frac{V_\xi(T_i)}{S_C(u)}dM_i^C(u) + o_p(n^{-1/2}) \tag{2.12}$$

En cuanto a b_{n2}, se puede descomponer en tres partes como sigue

$$\begin{aligned} b_{n2} =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}\Big(S_C(X_i)S_C(X_j) - \widehat{K}_C(X_i)\widehat{K}_C(X_j)\Big)}{S_C(X_i)S_C(X_j)\widehat{K}_C(X_i)\widehat{K}_C(X_j)} \\ =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}}{\widehat{K}_C(X_i)\widehat{K}_C(X_j)} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}}{S_C(X_i)S_C(X_j)} \\ &- \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}}{S_C(X_j)\widehat{K}_C(X_i)} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}}{S_C(X_i)\widehat{K}_C(X_j)} \\ &+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}}{S_C(X_i)S_C(X_j)} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}}{S_C(X_j)\widehat{K}_C(X_i)} \\ &- \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}}{S_C(X_i)S_C(X_j)} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}}{S_C(X_i)\widehat{K}_C(X_j)} \end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}\Big(S_{C}(X_i)-\widehat{K}_{C}(X_i)\Big)\Big(S_{C}(X_j)-\widehat{K}_{C}(X_j)\Big)}{S_{C}(X_i)S_{C}(X_j)\widehat{K}_{C}(X_i)\widehat{K}_{C}(X_j)}\\
&+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}\Big(S_{C}(X_i)-\widehat{K}_{C}(X_i)\Big)}{S_{C}(X_i)S_{C}(X_j)\widehat{K}_{C}(X_i)}\\
&+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}\Big(S_{C}(X_j)-\widehat{K}_{C}(X_j)\Big)}{S_{C}(X_i)S_{C}(X_j)\widehat{K}_{C}(X_j)}\\
=&b_{n21}+b_{n22}+b_{n23}.
\end{aligned}
\tag{2.13}
$$

Ya vimos que para un tiempo dado x, $\frac{S_{C}(x)-\widehat{K}_{C}(x)}{S_{C}(x)} = \int_0^x \frac{\widehat{K}_{C}(u-)}{S_{C}(u)}\frac{dM^{C}(u)}{R(u)}$ converge en distribución a una variable normalmente distribuida: $\frac{S_{C}(X_i)-\widehat{K}_{C}(X_i)}{S_{C}(X_i)} = O_p(n^{-1/2})$ y $\frac{S_{C}(X_j)-\widehat{K}_{C}(X_j)}{S_{C}(X_j)} = O_p(n^{-1/2})$. Entonces

$$
\begin{aligned}
b_{n21} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}O_p(n^{-1/2})O_p(n^{-1/2})}{\widehat{K}_{C}(X_i)\widehat{K}_{C}(X_j)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}o_p(1)o_p(1)}{\widehat{K}_{C}(X_i)\widehat{K}_{C}(X_j)}\\
&= o_p(1)(\xi+o_p(n^{-1/2})) = o_p(n^{-1/2})\\
b_{n21} &= o_p(n^{-1/2})
\end{aligned}
$$

Por su parte

$$
\begin{aligned}
b_{n22} =&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\delta_i\delta_j V_{ij}\Big(S_{C}(X_i)-\widehat{K}_{C}(X_i)\Big)}{S_{C}(X_i)S_{C}(X_j)\widehat{K}_{C}(X_i)}\\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\Big(S_{C}(X_i)-\widehat{K}_{C}(X_i)\Big)}{S_{C}(X_i)\widehat{K}_{C}(X_i)}\frac{1}{n}\sum_{j=1}^{n}\frac{\delta_j V_{ij}}{S_{C}(X_j)}\\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\Big(S_{C}(X_i)-\widehat{K}_{C}(X_i)\Big)}{S_{C}(X_i)\widehat{K}_{C}(X_i)}\left(E\left(\frac{\delta_j V_{ij}}{S_{C}(X_j)}\middle| l_i\right)+o_p(1)\right)
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i V_\xi(T_i)\Big(S_{_C}(X_i)-\widehat{K}_{_C}(X_i)\Big)}{S_{_C}(X_i)\widehat{K}_{_C}(X_i)}\\
&+\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i O_{p_i}(n^{-1/2})}{\widehat{K}_{_C}(X_i)}\Big(o_p(1)\Big)\\
=&\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i V_\xi(T_i)\Big(S_{_C}(X_i)-\widehat{K}_{_C}(X_i)\Big)}{S_{_C}(X_i)\widehat{K}_{_C}(X_i)}+o_p(n^{-1/2})\\
=&\int_0^\infty\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i V_\xi(T_i)I(T_i\geq u)}{\widehat{K}_{_C}(X_i)\widehat{K}_T(u)}\frac{dM^{_C}(u)}{nS_{_C}(u)}+o_p(n^{-1/2})\\
=&\int_0^\infty\frac{\widehat{F}(V_\xi,u)}{nS_{_C}(u)}dM^{_C}(u)+o_p(n^{-1/2})\\
=&\frac{1}{n}\sum_{i=1}^{n}\int_0^\infty\frac{F(V_\xi,u)}{S_{_C}(u)}dM_i^{_C}(u)+o_p(n^{-1/2})
\end{aligned}
\tag{2.14}
$$

Sucede lo mismo con la tercera parte, es decir $b_{n23}=b_{n22}$. Reunamos los términos bn que representan $\widehat{\xi}-\xi$

$$
\begin{aligned}
\widehat{\xi}-\xi=&b_{n1}+b_{n21}+b_{n22}+b_{n23}\\
=&\frac{2}{n}\sum_{i=1}^{n}\Big(V_\xi(T_i)-\xi\Big)-\frac{2}{n}\sum_{i=1}^{n}\int_0^\infty\frac{V_\xi(T_i)}{S_{_C}(u)}dM_i^{_C}(u)\\
&+\frac{2}{n}\sum_{i=1}^{n}\int_0^\infty\frac{F(V_\xi,u)}{S_{_C}(u)}dM_i^{_C}(u)+o_p(n^{-1/2})
\end{aligned}
\tag{2.15}
$$

$\widehat{\mu}-\mu=O_p(n^{-1/2})$ y $\widehat{\xi}-\xi=O_p(n^{-1/2})$ por el teorema del límite central. Usando series de Taylor estocásticas para $\widehat{G}(\widehat{\mu},\widehat{\xi})$ alrededor de (μ,ξ)

$$
\begin{aligned}
\widehat{G}=&\frac{\xi}{\mu}-1+\frac{\widehat{\xi}-\xi}{\mu}-\frac{\xi}{\mu^2}(\widehat{\mu}-\mu)+o_p(n^{-1/2})\\
=&G+\frac{1}{\mu}\Big((\widehat{\xi}-\xi)-(G+1)(\widehat{\mu}-\mu)\Big)+o_p(n^{-1/2})
\end{aligned}
\tag{2.16}
$$

Sustituimos (2.9) y (2.15) en (2.16), restamos G y multiplicamos por $\sqrt{n}$

$$\begin{aligned}\sqrt{n}(\widehat{G}-G) =&\frac{1}{\mu}\Bigg[\Bigg(\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\Big(V_\xi(T_i)-\xi\Big)-\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \frac{V_\xi(T_i)}{S_{_C}(u)}dM_i^{_C}(u)\\ &+\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \frac{F(V_\xi,u)}{S_{_C}(u)}dM_i^{_C}(u)\Bigg)\\ &-(G+1)\Bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T_i-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \frac{T_i}{S_{_C}(u)}dM_i^{_C}(u)-\sqrt{n}\mu\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \frac{F(T,u)}{S_{_C}(u)}dM_i^{_C}(u)\Bigg)\Bigg]+o_p(1)\\ =&\frac{1}{\mu}\Bigg[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(2\Big(V_\xi(T_i)-\xi\Big)-(G+1)\Big(T_i-\mu\Big)\Bigg)\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty 2\Big(\frac{F(V_\xi,u)-V_\xi(T_i)}{S_{_C}(u)}\Big)\\ &-(G+1)\Big(\frac{F(T,u)-T_i}{S_{_C}(u)}\Big)dM_i^{_C}(u)\Bigg]+o_p(1)\end{aligned} \tag{2.17}$$

Aplicando el teorema del límite central tradicional y el teorema del límite central para martingalas a (2.17) probamos que $\sqrt{n}(\widehat{G}-G)$ converge a la distribución normal con media 0. Además

$$\begin{aligned}\Omega =&var\Bigg\{\frac{1}{\mu}\Bigg[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(2\Big(V_\xi(T_i)-\xi\Big)-(G+1)\Big(T_i-\mu\Big)\Bigg)\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty 2\Big(\frac{F(V_\xi,u)-V_\xi(T_i)}{S_{_C}(u)}\Big)\\ &-(G+1)\Big(\frac{F(T,u)-T_i}{S_{_C}(u)}\Big)dM_i^{_C}(u)\Bigg]+o_p(1)\Bigg\}\end{aligned}$$

(La varianza será igual a la suma de la varianzas de dos términos: la suma de martingalas y la suma de términos sin martingalas. Aplicando la definición de covarianza entre ambos términos puede verse que resulta en la esperanza del producto de ambos, utilizando la ley de la esperanza total obtenemos la esperanza de una martingala: cero.)

$$
\begin{aligned}
=&\frac{1}{\mu^2} var\Bigg\{ \frac{1}{\sqrt{n}}\Big(\sum_{i=1}^{n} 2V_\xi(T_i) - (G+1)T_i \Big) \\
&\qquad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\infty 2\Big(\frac{F(V_\xi, u) - V_\xi(T_i)}{S_{_C}(u)} \Big) \\
&\qquad\qquad - (G+1)\Big(\frac{F(T,u) - T_i}{S_{_C}(u)} \Big) dM_i^{_C}(u) \Bigg\} \\
=&\frac{1}{\mu^2} \Bigg\{ \frac{1}{n} var\Bigg[\sum_{i=1}^{n} \Big(2V_\xi(T_i) - (G+1)T_i \Big) \\
&\qquad + \sum_{i=1}^{n} \int_0^\infty 2\Big(\frac{F(V_\xi, u) - V_\xi(T_i)}{S_{_C}(u)} \Big) \\
&\qquad - (G+1)\Big(\frac{F(T,u) - T_i}{S_{_C}(u)} \Big) dM_i^{_C}(u) \Bigg] \Bigg\} \\
=&\frac{1}{\mu^2} \Bigg\{ var\Big[2V_\xi(T_i) - (G+1)T_i \Big] \\
&+ \frac{1}{n} var\Bigg[\sum_{i=1}^{n} \int_0^\infty 2\Big(\frac{F(V_\xi, u) - V_\xi(T_i)}{S_{_C}(u)} \Big) \\
&\qquad\qquad - (G+1)\Big(\frac{F(T,u) - T_i}{S_{_C}(u)} \Big) dM_i^{_C}(u) \Bigg] \Bigg\} \\
=&\frac{1}{\mu^2} \Bigg\{ var\Big[2V_\xi(T_i) - (G+1)T_i \Big] \\
&+ \frac{1}{n} E\Bigg[\sum_{i=1}^{n} \int_0^\infty \Big[2\Big(\frac{F(V_\xi, u) - V_\xi(T_i)}{S_{_C}(u)} \Big) \\
&\qquad\qquad - (G+1)\Big(\frac{F(T,u) - T_i}{S_{_C}(u)} \Big) \Big]^2 h_{_C}(u) R_i(u) du \Bigg] \Bigg\} \\
=&\frac{1}{\mu^2} \Bigg\{ var\Big[2V_\xi(T_i) - (G+1)T_i \Big] + E\Big[\int_0^\infty \Big[2\Big(\frac{F(V_\xi, u) - V_\xi(T_i)}{S_{_C}(u)} \Big) \\
&\qquad - (G+1)\Big(\frac{F(T,u) - T_i}{S_{_C}(u)} \Big) \Big]^2 h_{_C}(u) R_i(u) du \Big] \Bigg\}
\end{aligned}
$$

$$\begin{aligned}\Omega =&\frac{1}{\mu^2}\Bigg\{var[2V_\xi(T_i)-(G+1)T_i]+4\Bigg[E\Bigg(\int_0^\infty \frac{F(V_\xi,u)^2}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)\\
&+E\Bigg(\int_0^\infty \frac{V_\xi(T_i)^2}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)-2E\Bigg(\int_0^\infty \frac{F(V_\xi,u)V_\xi(T_i)}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)\Bigg]\\
&+(G+1)^2\Bigg[E\Bigg(\int_0^\infty \frac{F(T,u)^2}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)\\
&+E\Bigg(\int_0^\infty \frac{T_i^2}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)-2E\Bigg(\int_0^\infty \frac{F(T,u)T_i}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)\Bigg]\\
&-4(G+1)\Bigg[E\Bigg(\int_0^\infty \frac{F(T,u)F(V_\xi,u)}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)\\
&+E\Bigg(\int_0^\infty \frac{T_iV_\xi(T_i)}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)-E\Bigg(\int_0^\infty \frac{F(T,u)V_\xi(T_i)}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)\\
&-E\Bigg(\int_0^\infty \frac{F(V_\xi,u)T_i}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Bigg)\Bigg]\Bigg\}\end{aligned} \tag{2.18}$$

con lo queda demostrado el primer teorema.

La realización de inferencias con este estimador requiere de la estimación de la varianza, por ello definamos los siguientes estimadores

$$\widehat{F}(T,x)=\frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_jX_jI(X_j\geq x)}{\widehat{K}_{_C}(X_j)}=\frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_jT_jI(T_j\geq x)}{\widehat{K}_{_C}(X_j)}$$

$$\widehat{F}(T^2,x)=\frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_jT_j^2I(T_j\geq x)}{\widehat{K}_{_C}(X_j)}$$

$$\widehat{F}(\widehat{V}_\xi,x)=\frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_j\widehat{V}_\xi(T_j)I(T_j\geq x)}{\widehat{K}_{_C}(X_j)}$$

$$\widehat{F}(\widehat{V}_\xi^2,x)=\frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_j\widehat{V}_\xi(T_j)^2I(T_j\geq x)}{\widehat{K}_{_C}(X_j)}$$

$$\widehat{F}(\widehat{V}_\xi\widehat{F}(\widehat{V}_\xi,x),x)=\frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_j\widehat{V}_\xi(T_j)\widehat{F}(\widehat{V}_\xi,x)I(T_j\geq x)}{\widehat{K}_{_C}(X_j)}$$

$$\begin{aligned}
\widehat{F}(T\widehat{F}(T,x),x) =& \frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_j T_j \widehat{F}(T,x) I(T_j \geq x)}{\widehat{K}_C(X_j)} \\
\widehat{F}(T\widehat{V}_\xi,x) =& \frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_j T_j \widehat{V}_\xi(T_j) I(T_j \geq x)}{\widehat{K}_C(X_j)} \\
\widehat{F}(T\widehat{F}(\widehat{V}_\xi,x),x) =& \frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_j T_j \widehat{F}(\widehat{V}_\xi,x) I(T_j \geq x)}{\widehat{K}_C(X_j)} \\
\widehat{F}(\widehat{V}_\xi\widehat{F}(T,x),x) =& \frac{1}{n\widehat{K}_T(x)}\sum_{j=1}^{n}\frac{\delta_j \widehat{V}_\xi(T_j) \widehat{F}(T,x) I(T_j \geq x)}{\widehat{K}_C(X_j)}
\end{aligned}$$

con

$$\widehat{V}_\xi(x) = xF_{nc}(x) + \frac{1}{n}\sum_{k=1}^{n}\frac{\delta_k T_k I(T_k \geq x)}{\widehat{K}_C(X_k)}.$$

Entonces

$$\begin{aligned}
\widehat{\Omega} =& \frac{1}{\widehat{\mu}^2}\Bigg\{\widehat{var}[2V_\xi(T_i) - (\widehat{G}+1)T_i] + 4\Bigg[\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(\widehat{V}_\xi,X_i)^2(1-\delta_i)}{\widehat{K}_C(X_i)^2} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(\widehat{V}_\xi^2,X_i)(1-\delta_i)}{\widehat{K}_C(X_i)^2} - \frac{2}{n}\sum_{i=1}^{n}\frac{\widehat{F}(\widehat{V}_\xi\widehat{F}(\widehat{V}_\xi,X_i),X_i)(1-\delta_i)}{\widehat{K}_C(X_i)^2}\Bigg] \\
&+ (\widehat{G}+1)^2\Bigg[\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(T,X_i)^2(1-\delta_i)}{\widehat{K}_C(X_i)^2} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(T^2,X_i)(1-\delta_i)}{\widehat{K}_C(X_i)^2} - \frac{2}{n}\sum_{i=1}^{n}\frac{\widehat{F}(T\widehat{F}(T,X_i),X_i)(1-\delta_i)}{\widehat{K}_C(X_i)^2}\Bigg] \\
&- 4(\widehat{G}+1)\Bigg[\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(T,X_i)\widehat{F}(\widehat{V}_\xi,X_i)(1-\delta_i)}{\widehat{K}_C(X_i)^2} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(T\widehat{V}_\xi,X_i)(1-\delta_i)}{\widehat{K}_C(X_i)^2} - \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(\widehat{V}_\xi\widehat{F}(T,X_i),X_i)(1-\delta_i)}{\widehat{K}_C(X_i)^2} \\
&- \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(T\widehat{F}(\widehat{V}_\xi,X_i),X_i)(1-\delta_i)}{\widehat{K}_C(X_i)^2}\Bigg]\Bigg\} \\
=& \widehat{\Omega}_1 + \cdots + \widehat{\Omega}_{11}
\end{aligned}$$

Teorema 2. *Cuando se cumplen las suposiciones 1 a 3, a medida que $n \to \infty$, se cumple que*

$$\widehat{\Omega} = \Omega + o_p(1)$$

donde

$$\widehat{var}[2V_\xi(T_i) - (\widehat{G}+1)T_i]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \frac{\delta_i\Big[2\Big(\widehat{V}_\xi(T_i) - \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i \widehat{V}_\xi(T_i)}{\widehat{K}_{C}(X_i)}\Big) - (\widehat{G}+1)\Big(T_i - \widehat{\mu}\Big)\Big]^2}{\widehat{K}_{C}(X_i)}$$

Demostración:
Primero, veamos que

$$\begin{aligned} &var[2V_\xi(T_i) - (G+1)T_i] \\ &\quad = E\Big[\Big((2V_\xi(T_i) - (G+1)T_i) - E(2V_\xi(T_i) - (G+1)T_i)\Big)^2\Big] \\ &\quad = E\Big[\Big(2(V_\xi(T_i) - E(V_\xi(T_i))) - (G+1)(T_i - \mu)\Big)^2\Big]. \end{aligned}$$

Llamemos $\widehat{var}[2V_\xi(T_i) - (\widehat{G}+1)T_i]$ al estimador $\widehat{var}[2V_\xi(T_i) - (\widehat{G}+1)T_i]$ pero sustituyendo $\widehat{K}_{C}(x)$ con $S_{C}(x)$. Así que $\widehat{var}[2V_\xi(T_i) - (\widehat{G}+1)T_i] = \widehat{var}[2V_\xi(T_i) - (\widehat{G}+1)T_i] + o_p(1)$.

Luego, como en la primera parte del primer teorema, vimos que aplicando la ley de los grandes números y la ley de la esperanza total $\widehat{\mu} = \mu + o_p(1)$ y $\widehat{G} = G + o_p(1)$, lo mismo sucede con $\widehat{V}_\xi(T_i)$ y $\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i \widehat{V}_\xi(T_i)}{\widehat{K}_{C}(X_i)}$, son consistentes con sus respectivos parámetros. Tenemos que $\widehat{\Omega}_1 = \frac{1}{\widehat{\mu}^2}(\widehat{var}[2V_\xi(T_i) - (\widehat{G}+1)T_i]) = \Omega_1 + o_p(1) = var[2V_\xi(T_i) - (G+1)T_i] + o_p(1)$.

Con respecto a los restantes 10 términos de la varianza, notemos que incluyen integrales respecto del proceso de intensidad $h_{C}(u)R_i(u)$, por ello, de manera análoga a los estimadores Nelson-Aalen y Kaplan-Meier, sustituimos $h_{C}(u)du$ con $\frac{dN_i^{C}(u)}{R_i(u)}$ para tener estimadores consistentes.

Tomando como ejemplo $\Omega_2 = \frac{4}{\mu^2}E\Big(\int_0^\infty \frac{F(V_\xi,u)^2}{S_{C}(u)^2}h_{C}(u)R_i(u)du\Big)$ procedemos como

sigue

$$\frac{4}{\mu^2}E\Big(\int_0^\infty \frac{F(V_\xi,u)^2}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\Big) \approx \frac{4}{\mu^2}E\Big(\int_0^\infty \frac{F(V_\xi,u)^2}{S_{_C}(u)^2}dN_i^{_C}(u)\Big)$$
$$=\frac{4}{\mu^2}E\Big(\sum_{j=1}^{n}\frac{F(V_\xi,X_j)^2}{S_{_C}(X_j)^2}dN_i^{_C}(X_j)\Big)$$

(la suma $\sum_{j=1}^{n}\frac{F(V_\xi,X_j)^2}{S_{_C}(X_j)^2}dN_i^{_C}(X_j)$ solo puede tener un elemento diferente de 0 e igual a $\frac{F(V_\xi,X_i)^2}{S_{_C}(X_i)^2}$ debido a que $dN_i^{_C}(X_j) = I(X_i = X_j, \delta_i = 0)$)

$$\approx\frac{4}{\widehat{\mu}^2}E\Big(\frac{\widehat{F}(\widehat{V}_\xi,X_i)^2}{\widehat{K}_{_C}(X_i)^2}(1-\delta_i)\Big)$$
$$\approx\frac{4}{\widehat{\mu}^2}\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{F}(\widehat{V}_\xi,X_i)^2(1-\delta_i)}{\widehat{K}_{_C}(X_i)^2}\Big)$$

que es precisamente $\widehat{\Omega}_2$. Además, $\widehat{\Omega}_2 = \Omega_2 + o_p(1)$ razonando como antes, por la consistencia de $\widehat{\mu}$, $\widehat{K}_{_C}(x)$, $\widehat{F}(\widehat{V}_\xi, X_i)$ y la consistencia de $\int_0^\infty \frac{F(V_\xi,u)^2}{S_{_C}(u)^2}dN_i^{_C}(u)$ con $\int_0^\infty \frac{F(V_\xi,u)^2}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du$, utilizando la ley de los grandes números y la ley de la esperanza total.

Este razonamiento se repite para los últimos 9 términos, pero falta un detalle y es que al observar solamente (X, δ) no podemos utilizar el estimador $\widehat{V}_\xi(T_i)$ ni T_i como lo sugieren los términos $\Omega_3, \Omega_4, \Omega_6, \Omega_7, \Omega_9, \Omega_{10}, \Omega_{11}$, así que debemos dar un rodeo. Tomemos como ejemplo Ω_4 y veamos que

$$
\begin{aligned}
&2E\left(\int_0^\infty \frac{F(V_\xi,u)V_\xi(T_i)}{S_{_C}(u)^2}h_{_C}(u)R_i(u)du\right)\\
&\quad=2E\left(\int_0^\infty \frac{F(V_\xi,u)V_\xi(T_i)I(T_i>u)I(C_i>u)}{S_{_C}(u)^2}h_{_C}(u)du\right)\\
&\quad=2E\left(\int_0^\infty \frac{E\Big(F(V_\xi,u)V_\xi(T_i)I(T_i>u)\Big)I(C_i>u)}{S_{_C}(u)^2}h_{_C}(u)du\right)\\
&\quad=2E\left(\int_0^\infty \frac{F\Big(F(V_\xi,u)V_\xi,u\Big)S_T(u)I(C_i>u)}{S_{_C}(u)^2}h_{_C}(u)du\right)\\
&\quad=2E\left(\int_0^\infty \frac{F\Big(F(V_\xi,u)V_\xi,u\Big)}{S_{_C}(u)^2}R_i(u)h_{_C}(u)du\right).
\end{aligned}
$$

La misma lógica se aplica a los demás términos mencionados y es así como obtenemos los estimadores de los términos de $\widehat{\Omega}$, completando la prueba del segundo teorema.

2.3. Estimación con censura dependiente de covariables

En casos reales es poco probable que haya independencia entre T_i y C_i. Un buen ejemplo para entender esta falta de independencia entre C_i y T_i se encuentra en Altman 1991, sección 13.2.1. Esta sección describe un experimento medico relacionado con el mareo en el que un grupo de personas fueron puestas en una cabina movida verticalmente por una máquina. Las personas podían experimentar el movimiento en la cabina hasta por dos horas, el evento de interés era el momento del vómito de cada participante. Algunos participantes pidieron detener la máquina y salir de la cabina a pesar de no haber vomitado, dando como resultado tiempos de censura, mientras otros tantos sí pudieron resistir las dos horas sin vomitar, siendo también censurados. Este experimento muestra como el mareo influye tanto en el tiempo de censura como en el tiempo hasta el vómito, ya que es más probable que una persona que se siente mareada solicite bajar de la cabina o vomite, que una persona que no siente mareo.

Suponemos una muestra $(X_i, \delta_i, \boldsymbol{Z}_i)$ de variables *i.i.d* donde $X_i = min(T_i, C_i)$, $\delta_i = 1$ si $X_i = T_i$ y $\boldsymbol{Z}_i$ un vector de covariables. Denotemos $F_C^Z(x) = P(C \leq x|Z)$, luego $S_C^Z(x) = 1 - F_C^Z(x)$, $H_C^Z(x) = -\int_0^x \frac{dS_C^Z(u)}{S_C^Z(u)}$ y $h_C^Z(x) = \frac{dH_C^Z(x)}{dx}$.

C_i y T_i son variables condicionalmente independientes dado $\boldsymbol{Z}_i$. Dos eventos, A y B, son condicionalmente independientes dado el evento C si las probabilidades condicionadas de A y B respecto de C son independientes: $A \perp\!\!\!\perp B|Z \iff P(A \cap B|C) = P(A|C)P(B|C)$, equivalentemente $A \perp\!\!\!\perp B|Z \iff P(A|B \cap C) = P(A|C)$. En términos de vectores de variables aleatorias: $C \perp\!\!\!\perp T|Z \iff F_{X,Y|Z}(x,y) = F_{X|Z}(x)F_{Y|Z}(y)$. [2]

Similar a la sección anterior

$$E(\delta|T,\boldsymbol{Z}) = P(\delta = 1|T,\boldsymbol{Z}) = P(C > T|T,\boldsymbol{Z}) = S_C^Z(T),$$

entonces

$$E\left[\frac{g(X)\delta}{S_C^Z(X)}\right] = E\left[E\left(\frac{g(X)\delta}{S_C^Z(X)}\middle|T,\boldsymbol{Z}\right)\right] = E\left[\frac{g(T)E(\delta|T,\boldsymbol{Z})}{S_C^Z(T)}\right] = E(g(T)).$$

La independencia entre C_i y T_i implica que la variable de censura puede depender solo de información del "pasado", no del futuro. En otras palabras, la variable de censura puede depender solo de covariables incluidas en el modelo. El estimador Kaplan-Meier convencional es un modelo sin covariables, cuando la censura depende de covariables el estimador Kaplan-Meier puede resultar sesgado. (O. Aalen, Borgan y Gjessing 2008, p. 175)

Para obtener de forma sencilla un estimador de $S_C^Z(X)$ utilizamos el modelo aditivo de Aalen para censura dependiente de covariables como en Satten, Datta y J. Robins 2001:

$$h_C^{Z_i}(x) = \boldsymbol{Z}_i^T \boldsymbol{\beta}(x)$$

donde el primer elemento de cada vector $\boldsymbol{Z}_i$ es un 1.

Estimamos $H_C(x|\boldsymbol{Z}_i) = \int_0^x h_C^{Z_i}(u)du = \int_0^x \boldsymbol{Z}_i^T d\boldsymbol{B}(u)$. $\boldsymbol{B}(x)$ se estima por

$$\widehat{\boldsymbol{B}}(x) = \sum_{i=1}^{n} I(X_i \leq x)(1-\delta_i)\boldsymbol{A}^{-1}(X_i)\boldsymbol{Z}_i,$$

donde

$$\boldsymbol{A}(x) = \sum_{i=1}^{n} I(X_i \geq x)\boldsymbol{Z}_i\boldsymbol{Z}_i^T.$$

[2] Información revisada en: Colaboradores de Wikipedia. (2019, 28 de marzo). Independencia condicional. En Wikipedia, La Enciclopedia Libre . Consultado el 22 de abril de 2019, de https://en.wikipedia.org/w/index.php?title=Conditional_independence&oldid=889885084.

Cuando $\boldsymbol{A}(x)$ no sea invertible ocuparemos la *inversa espectral* $P \cdot Diag(E^{+}) \cdot P^{T}$. De esta manera calcularemos los estimadores de las curvas de supervivencia dado $\boldsymbol{Z}_i$ como

$$\widehat{K}_C^{Z_i}(x) = \prod_{X_j \leq x} (1 - d\widehat{H}_C^{Z_i}(X_j)),$$

Después de obtener todos los estimadores $\{\widehat{K}_C^{Z_i}(X_i)\}_i^n$ para nuestra muestra, podemos estimar el nuevo estimador del coeficiente de Gini para censura dependiente de covariables

$$\widehat{G}_Z = \frac{\widehat{\xi}_Z}{\widehat{\mu}_Z} - 1 \tag{2.19}$$

donde $\widehat{\xi} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i 2F_{nc}^{Z}(X_i)X_i}{\widehat{K}_C^{Z_i}(X_i)}$, $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i X_i}{\widehat{K}_C^{Z_i}(X_i)}$ y $F_{nc}^{Z}(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j I(X_j \leq x)}{\widehat{K}_C^{Z_i}(X_j)}$.

Al igual que el estimador del coeficiente de Gini bajo censura independiente, $\widehat{G}_Z$ cumple algunas propiedades asintóticas. Sea $\tilde{\boldsymbol{Z}}_i(x) = I(X_i \geq x)\boldsymbol{Z_i}$, $\tilde{\mathcal{Z}}(x) = (\tilde{\boldsymbol{Z}}_1(x)), \ldots, \tilde{\boldsymbol{Z}}_n(x))$ y nombremos

$$\Upsilon^{T}(x) = E\left[\frac{\delta_i\Big(2V_\xi(T_i) - (G+1)T_i\Big)I(T_i > x)\boldsymbol{Z}_i^{T}}{\widehat{K}_C^{Z\,i}(x)}\right],$$

$$\sigma^2 = plim_{n\to\infty} n^{-1}\int_0^{\infty} \boldsymbol{\tau}_n^{T}(x) diag\{I(X_1 \geq x)h_C^{Z_1}(x), \ldots, I(X_n \geq x)h_C^{Z_n}(x)\}\boldsymbol{\tau}_n(x)dx,$$

donde $\boldsymbol{\tau}_n^{T}(x) = \xi^{T}(x) - \Upsilon^{T}(x)\boldsymbol{a}(x)\tilde{\mathcal{Z}}(x)$,

$$\boldsymbol{a}(x) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i(x)\tilde{\boldsymbol{Z}}_i^{T}(x) = E[\tilde{\boldsymbol{Z}}_i(x)\tilde{\boldsymbol{Z}}_i^{T}(x)],$$

(como en Lv, G. Zhang, Li y col. 2016), y

$$\xi^{T}(x) = \left(\frac{2V_\xi(T_1) - (G+1)T_1}{\widehat{K}_C^{Z_1}(x)}, \cdots, \frac{2V_\xi(T_n) - (G+1)T_n}{\widehat{K}_C^{Z_n}(x)}\right)$$

Nuevamente hay tres condiciones que sostienen las propiedades de consistencia y convergencia en distribución del estimador $\widehat{G}_Z$.

Suposición 1 La muestra $\{X_i, \boldsymbol{Z}_i, \delta_i\}_{i=1}^{n}$ está compuesta de variables independientes e idénticamente distribuidas (v.i.i.d.) y las variables T y C provienen de distribuciones que son continuas y positivas.

Suposición 2 $\boldsymbol{Z}_i$ puede explicar toda la dependencia entre las variables C_i y T_i, hay una independencia condicionada a $\boldsymbol{Z}_i$, $C_i \perp\!\!\!\perp T_i|\boldsymbol{Z}_i$ y la tasa de riesgo con respecto a C_i es $h_C^{Z_i}(x) = \boldsymbol{Z}_i^T \boldsymbol{\beta}(x)$.

Suposición 3 $Var[2V_\xi(T_i) - (G+1)T_i]$ y σ^2 existen.

El siguiente teorema muestra que el estimador $\widehat{G}_Z$ es consistente y asintóticamente normal.

Teorema 3. *1. Cuando se cumplen las suposiciones 1 y 2, a medida que* $n \to \infty$, *sucede que* $\widehat{G}_Z = G + o_p(1)$. *2. Cuando se cumplen las 3 suposiciones, a medida que* $n \to \infty$, *se tiene que*

$$\sqrt{n}(\widehat{G}_Z - G) \xrightarrow{d} N(0, \Omega_Z),$$

donde

$$\Omega_Z = \frac{1}{\mu^2}\left(Var[2V_\xi(T_i) - (G+1)T_i] + \sigma^2\right).$$

Demostración:

1. En el primer capítulo mostramos entre las propiedades del estimador $\widehat{K}_C^{Z_i}(x)$ que es consistente, es decir $\widehat{K}_C^{Z_i}(x) = S_C^{Z_i}(x) + o_p(1)$. El resto de la prueba de esta primera parte se resuelve de manera indirecta como en la primera parte de la demostración del primer teorema (ver primera parte del **Teorema 1**).

2. Verifiquemos ahora la normalidad del estimador y su varianza. Comencemos aproximando la diferencia $\widehat{\mu}_Z - \mu$ y, de manera análoga a (2.4), (2.5) y (2.6), veamos que

$$\begin{aligned}\widehat{\mu}_Z - \mu =& \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i}{\widehat{K}_C^{Z_i}(X_i)} - \mu \\ =& \frac{1}{n}\sum_{i=1}^{n} T_i - \mu - \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty \frac{T_i}{S_C^{Z_i}(u)} dM_i^C(u) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i\left(S_C^{Z_i}(X_i) - \widehat{K}_C^{Z_i}(X_i)\right)}{S_C^{Z_i}(X_i)\widehat{K}_C^{Z_i}(X_i)}.\end{aligned} \tag{2.20}$$

Como

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i\Big(S_C^{Z_i}(X_i)-\widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)\widehat{K}_C^{Z_i}(X_i)}-\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i\Big(S_C^{Z_i}(X_i)-\widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)^2}\\ =\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i\Big(S_C^{Z_i}(X_i)-\widehat{K}_C^{Z_i}(X_i)\Big)^2}{S_C^{Z_i}(X_i)^2\widehat{K}_C^{Z_i}(X_i)}=o_p(n^{-1/2}),\end{aligned} \tag{2.21}$$

(2.20) se puede reexpresar como

$$\begin{aligned}\widehat{\mu}_Z-\mu=&\frac{1}{n}\sum_{i=1}^{n}T_i-\mu-\frac{1}{n}\sum_{i=1}^{n}\int_0^\infty\frac{T_i}{S_C^{Z_i}(u)}dM_i^C(u)\\ &+\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i T_i\Big(S_C^{Z_i}(X_i)-\widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)^2}+o_p(n^{-1/2}).\end{aligned} \tag{2.22}$$

Luego, operando de manera análoga a (2.10), (2.11), (2.12), (2.13) y (2.14), tenemos para

$$\begin{aligned}\widehat{\xi}_Z-\xi=&\frac{2}{n}\sum_{i=1}^{n}\Big(V_\xi(T_i)-\xi\Big)-\frac{2}{n}\sum_{i=1}^{n}\int_0^\infty\frac{V_\xi(T_i)}{S_C^{Z_i}(u)}dM_i^C(u)\\ &+\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i V_\xi(T_i)\Big(S_C^{Z_i}(X_i)-\widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)\widehat{K}_C^{Z_i}(X_i)}+o_p(n^{-1/2}),\end{aligned} \tag{2.23}$$

con la diferencia de que en el paso análogo a (2.14) no aplicamos la sustitución de la representación martingala.

Así como en (2.21)

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i V_\xi(T_i)\Big(S_C^{Z_i}(X_i)-\widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)\widehat{K}_C^{Z_i}(X_i)}\\ &\quad=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i V_\xi(T_i)\Big(S_C^{Z_i}(X_i)-\widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)^2}+o_p(n^{-1/2}),\end{aligned} \tag{2.24}$$

de modo que

$$\begin{aligned}\widehat{\xi}_Z - \xi =& \frac{2}{n}\sum_{i=1}^{n}\Big(V_\xi(T_i) - \xi\Big) - \frac{2}{n}\sum_{i=1}^{n}\int_0^\infty \frac{V_\xi(T_i)}{S_C^{Z_i}(u)} dM_i^C(u) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i V_\xi(T_i)\Big(S_C^{Z_i}(X_i) - \widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)^2} + o_p(n^{-1/2}).\end{aligned} \tag{2.25}$$

Similar a (2.16), usando series de Taylor estocásticas para $\widehat{G}_Z(\widehat{\mu}_Z, \widehat{\xi}_Z)$ alrededor de (μ, ξ)

$$\begin{aligned}\widehat{G}_Z =& \frac{\xi}{\mu} - 1 + \frac{\widehat{\xi}_Z - \xi}{\mu} - \frac{\xi}{\mu^2}(\widehat{\mu}_Z - \mu) + o_p(n^{-1/2}) \\ =& G + \frac{1}{\mu}\Big((\widehat{\xi}_Z - \xi) - (G+1)(\widehat{\mu}_Z - \mu)\Big) + o_p(n^{-1/2}).\end{aligned} \tag{2.26}$$

Sustituimos (2.22) y (2.25) en (2.26), restamos G y multiplicamos por $\sqrt{n}$

$$\begin{aligned}\sqrt{n}(\widehat{G}_Z - G) =& \frac{1}{\mu}\Bigg[\Bigg(\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\Big(V_\xi(T_i) - \xi\Big) - \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \frac{V_\xi(T_i)}{S_C^{Z_i}(u)} dM_i^C(u) \\ &+ \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i V_\xi(T_i)\Big(S_C^{Z_i}(X_i) - \widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)^2}\Bigg) \\ &- (G+1)\Bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} T_i - \sqrt{n}\mu - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \frac{T_i}{S_C^{Z_i}(u)} dM_i^C(u) \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i T_i\Big(S_C^{Z_i}(X_i) - \widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)^2}\Bigg)\Bigg] + o_p(1) \\ =& \frac{1}{\mu}\Bigg[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(2\Big(V_\xi(T_i) - \xi\Big) - (G+1)\Big(T_i - \mu\Big)\Big) \\ &- \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \frac{2V_\xi(T_i) - (G+1)T_i}{S_C^{Z_i}(u)} dM_i^C(u) \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i\Big(2V_\xi(T_i) - (G+1)T_i\Big)\Big(S_C^{Z_i}(X_i) - \widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)^2}\Bigg] \\ &+ o_p(1)\end{aligned} \tag{2.27}$$

Si denotamos $Q_i = 2V_\xi(T_i) - (G+1)T_i$ y de forma similar a (1.14), utilizando el hecho de que $\frac{\widehat{K}_C^{Z_i}(u-)}{S_C^{Z_i}(u)} = 1 + o_p(1)$, podemos ver que

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i\Big(2V_\xi(T_i)-(G+1)T_i\Big)\Big(S_C^{Z_i}(X_i)-\widehat{K}_C^{Z_i}(X_i)\Big)}{S_C^{Z_i}(X_i)^2}\\
&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i Q_i}{S_C^{Z_i}(X_i)}\int_0^{X_i-}\boldsymbol{Z}_i^T\boldsymbol{A}^{-1}(u)\tilde{\mathcal{Z}}(u)d\boldsymbol{M}^C(u)+o_p(1)\\
&\quad=\frac{1}{\sqrt{n}}\int_0^{\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Q_i}{S_C^{Z_i}(X_i)}I(X_i>u)\boldsymbol{Z}_i^T(n^{-1}\boldsymbol{A})^{-1}(u)\tilde{\mathcal{Z}}(u)d\boldsymbol{M}^C(u)+o_p(1)\\
&\quad=\frac{1}{\sqrt{n}}\int_0^{\infty}\Upsilon^T(u)(\boldsymbol{a}(u))^{-1}\tilde{\mathcal{Z}}(u)d\boldsymbol{M}^C(u)+o_p(1),
\end{aligned}\tag{2.28}$$

Por último, sustituimos (2.28) en (2.27)

$$\begin{aligned}
\sqrt{n}(\widehat{G}_Z-G)=&\frac{1}{\mu}\Bigg[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(2\Big(V_\xi(T_i)-\xi\Big)-(G+1)\Big(T_i-\mu\Big)\Big)\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\infty}\frac{2V_\xi(T_i)-(G+1)T_i}{S_C^{Z_i}(u)}dM_i^C(u)\\
&+\frac{1}{\sqrt{n}}\int_0^{\infty}\Upsilon^T(u)(\boldsymbol{a}(u))^{-1}\tilde{\mathcal{Z}}(u)d\boldsymbol{M}^C(u)\Bigg]\\
&+o_p(1)
\end{aligned}\tag{2.29}$$

tomando en consideración que el integrando del último término es localmente acotado y predecible (Andersen y col. 1993, proposición II.4.1) para aplicar teorema central de límite para martingalas, probando la normalidad asintótica del estimador $\widehat{G}_Z$.

La estimación directa de $\Omega_Z = \frac{1}{\mu^2}\big(Var[2V_\xi(T_i)-(G+1)T_i]+\sigma^2\big)$ es más difícil que en el caso de censura independiente, por ello es recomendable utilizar métodos *bootstrap* para su aproximación.

Capítulo 3

Simulación y aplicación a datos empíricos

3.1. Simulación

3.1.1. Simulación para censura independiente

El experimento de censura independiente lo implementaremos en el programa estadístico `R`, de la siguiente manera. Generamos T y C de manera independiente, nosotros decidimos qué distribución siguen y con qué parámetros, la única condición que queremos satisfacer es que el porcentaje de datos censurados en cada muestra sea aproximadamente $50-55\,\%$. Tomamos muestras de $X_i = mín(T_i, C_i)$ y generamos su correspondiente indicador de censura, δ_i.

Con la idea de hacer un experimento más realista generamos primero la población, de tamaño 10^7, de la que calculamos su coeficiente de Gini real. Simulamos 3 poblaciones distintas, repetimos 500 veces: extraer muestras aleatorias simples de tamaño 100, 200 y 300, y calcular los estimadores.

Comparamos los resultados del estimador estudiado con otros estimadores del coeficiente de Gini para evaluar su conveniencia. Los estimadores analizados en este trabajo se basan en ponderar el estimador de Qin, Rao y Wu 2010.

Los otros estimadores utilizados para la comparación serán el de Bonetti, Gigliarano y Muliere 2009, del cual nos apoyamos con su propio paquete de `R`, `Survgini`(Gigliarano y Bonetti 2011), para los cálculos del coeficiente y su varianza.

Calcularemos el coeficiente ordinario, ignorando que algunos datos son C_i y tomando las X_i de manera homogénea, este cálculo se puede llevar a cabo con cualquier estimador simple del coeficiente de Gini, nosotros nos apoyamos en el paquete `laeken`(Alfons y Templ 2013).

El estimador estudiado en este trabajo suele sobre estimar el verdadero valor para muestras pequeñas, Davidson 2009 indica que un estimador construido a partir de (1.3) puede ser ambiguo pues su resultado depende de si se asume que la distribución es continua positiva o continua negativa. Un estimador que disminuye este problema es

$$\widehat{G}_w = \frac{2\sum_{i=1}^{n}(w_i X_i \sum_{j=1}^{n} w_j) - \sum_{i=1}^{n} w_i^2 X_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i X_i} - 1, \tag{3.1}$$

cuyas w_i pueden ser sustituidas por los ponderadores $\frac{\delta_i}{\widehat{K}_C(X_i)}$ o $\frac{\delta_i}{\widehat{K}_C^{Z_i}(X_i)}$, para analizar censura independiente o dependiente de covariables, respectivamente. El paquete laeken nos permite calcular este estimador cuando los ponderadores son $\frac{\delta_i}{\widehat{K}_C^{Z_i}(X_i)}$.

A pesar de generar T y C de manera independiente, probamos los estimadores para censura dependiente de covariables con la intención de comparar su desempeño en un caso donde no son requeridos. Generamos una covariable de manera independiente, $Y \sim weibull(2, 20)$, donde $weibull(a, b)$ es una distribución weibull con parámetro de forma a y parámetro de escala b.

Los 3 conjuntos de datos que generamos son los siguientes:

1. $D1$: $T \sim \chi^2(20)$, $C \sim Exp(0.035)$ y $X = mín(T, C)$
2. $D2$: $T \sim Exp(0.3)$, $C \sim F(2, 20)$ y $X = mín(T, C)$
3. $D3$: $T \sim F(3, 3)$, $C \sim Exp(1)$ y $X = mín(T, C)$

donde $Exp(\lambda)$ es la distribución exponencial con parámetro de tasa $\lambda = 1/media$, $F(gl_1, gl_2)$ es la distribución *Fisher–Snedecor* con gl_1, gl_2 grados de libertad. Los correspondientes valores reales del coeficiente de Gini son: 0.1761802, 0.5001255 y 0.7232401, respectivamente.

Reportamos los resultados de la simulación en una tabla, indicamos el error medio, el error cuadrado medio (que multiplicamos por 100 por ser originalmente muy pequeño) y las probabilidades de cobertura que indican el porcentaje de las simulaciones en las que el verdadero valor cayó dentro de los intervalos de confianza al 95 %.

En los casos de dependencia, del estimador ordinario y del estimador $\widehat{G}_w$ los intervalos proceden de esquemas de remuestreo bootstrap de 200 repeticiones y obtenemos los cuantiles correspondientes al 95 %. La función `gini` del paquete `laeken` arroja también la varianza por remuestreo bootstrap si así lo requerimos. En el caso de $\widehat{G}_Z$, en cada una de las 500 repeticiones del experimento, extraemos 200 submuestras sin remplazo de tamaño $(n \cdot 0.9)$ para generar los cálculos y poder determinar un intervalo.

Tabla 3.1: Resultados de simulación para censura independiente

n	Estimador	Error	ECM(x10^2)	P.C.	Error	ECM(x10^2)	P.C.
		D1			D2		
100	Bon	-0.005	0.031	0.118	-0.047	0.512	0.07
	Ord.	0.140	2.028	0	-0.008	0.128	0.968
	Ind.	0.018	0.066	0.844	0.024	0.507	0.856
	Alt.Ind.	-0.005	0.031	1	-0.047	0.512	0.652
	Dep.	0.007	0.078	0.502	-0.062	1.317	0.25
	Alt.Dep.	-0.006	0.034	1	-0.048	0.494	0.662
200	Bon	-0.002	0.017	0.094	-0.026	0.222	0.05
	Ord.	0.143	2.077	0	0.002	0.059	0.86
	Ind.	0.009	0.026	0.89	0.011	0.214	0.874
	Alt.Ind.	-0.002	0.017	0.952	-0.026	0.222	0.538
	Dep.	0.003	0.029	0.502	-0.036	0.506	0.262
	Alt.Dep.	-0.003	0.016	0.955	-0.030	0.240	0.494
300	Bon	-0.002	0.010	0.062	-0.019	0.177	0.048
	Ord.	0.142	2.028	0	0.004	0.047	0.714
	Ind.	0.005	0.012	0.956	0.005	0.174	0.862
	Alt.Ind.	-0.002	0.010	0.914	-0.019	0.177	0.388
	Dep.	0.0009	0.014	0.502	-0.024	0.337	0.282
	Alt.Dep.	-0.003	0.010	0.913	-0.022	0.183	0.397
		D3					
100	Bon	-0.245	6.245	0			
	Ord.	-0.228	5.335	0			
	Ind.	-0.177	3.609	0.164			
	Alt.Ind.	-0.245	6.245	0			
	Dep.	-0.353	13.051	0			
	Alt.Dep.	-0.248	6.344	0			
200	Bon	-0.222	5.116	0			
	Ord.	-0.221	4.925	0			
	Ind.	-0.175	3.414	0.104			
	Alt.Ind.	-0.222	5.116	0			
	Dep.	-0.325	11.006	0.			
	Alt.Dep.	-0.232	5.546	0			

Tabla 3.1 continuación

500	Bon	-0.163	2.767	0
	Ord.	-0.204	4.171	0
	Ind.	-0.126	1.814	0.152
	Alt.Ind.	-0.163	2.767	0
	Dep.	-0.270	7.568	0
	Alt.Dep.	-0.187	3.577	0

De la tabla 3.1 podemos hacer algunas observaciones.

A medida que aumenta el tamaño de la muestra los estimadores, todos salvo el estimador ordinario, mejoran su precisión con muestras relativamente pequeñas. En el segundo conjunto de datos, el estimador ordinario se ve mejorar, pero esto es debido a que formulamos T y C muy similares, si uno se fijara en sus graficas de densidad ambas variables comparten mucho de su rango, pero esto obviamente no ocurrirá siempre en la práctica.

El estimador de Bonetti, Gigliarano y Muliere 2009 coincide con $\widehat{G}_w$ cuando $w_i = \delta_i/\widehat{K}_C(X_i)$. El estimador de Bonetti es bastante bueno pero su pequeña varianza hace que sus intervalos de confianza no contengan al verdadero valor.

Los estimadores formulados para censura dependiente de covariables son también consistentes y para algunos conjuntos de datos se comportan incluso mejor que los estimadores para censura independiente.

Cuando el coeficiente de Gini real es alto (solo remitiéndonos a la simulación diríamos que mayor a 0.5), todos los estimadores parecen subestimarlo por ello es que los resultados del conjunto 3 no son tan buenos como los del primer conjunto. A pesar de lo anterior, es en estos casos donde $\widehat{G}$ se muestra preferible por su mayor precisión según el error cuadrado medio. Podría decirse que el estimador de Lv, G. Zhang y Ren 2017 para censura independiente estudiado en este trabajo es preferible si se conoce a priori que el coeficiente real es alto, mayor a 0.5 y el porcentaje de datos censurados es importante.

En el caso de las probabilidades de cobertura vemos que los intervalos de confianza para el estimador de independencia son los que más se acercan a la proporción teórica del 95 %. Sin embargo, las probabilidades de cobertura del estimador para dependencia de covariables no parecen acercarse al valor teórico, lo que contrasta con los resultados de simulación presentados por Lv, G. Zhang y Ren 2017, a pesar de que manejamos proporciones similares de datos censurados. En nuestra opinión, nuestro experimento de simulación revela algunas debilidades del estimador analizado

y también de los demás que, a pesar de ser buenos por su consistencia, para algunos conjuntos de datos no resultan tan precisos con tamaños de muestra como los manejados en el artículo referido. Al parecer, la mayor variabilidad de los datos provoca un difícil cálculo haciendo que los intervalos fallen en incluir el valor verdadero.

3.1.2. Simulación para censura dependiente de covariables

El experimento de censura dependiente de covariables los haremos de forma muy similar al caso de la censura independiente. Aquí comenzamos por generar un covariable que será siempre una variable binomial, $Y \sim bin(n, 0,5)$, o sea que cada elemento de esta variable es un experimento *Bernoulli* con probabilidad de éxito 0.5. A partir de ella generamos T y C de manera independiente para cumplir la condición de independencia condicional.

De igual manera nosotros decidimos qué distribución ocupamos y con qué parámetros, buscando satisfacer que el porcentaje de datos censurados en cada muestra sea aproximadamente $50 - 55\,\%$. Generamos primero la población, de tamaño 10^7, de la que calculamos su coeficiente de Gini real. Simulamos 3 poblaciones distintas y repetimos 500 veces: extraer muestras aleatorias simples de tamaño 300, 600 y 1200, y calcular los estimadores.

Comparamos los resultados del estimador estudiado con los mismos de la subsección anterior, incluimos los estimadores de censura independiente. Esos estimadores se calculan de la misma manera que en el experimento previo.

Los 3 conjuntos de datos que generamos son los siguientes:

1. $D1$: $T = Y \cdot weibull(2.3, 2) + (1 - Y) \cdot weibull(1.5, 5)$,
   $C = Y \cdot weibull(2.3, 3) + (1 - Y) \cdot lognorm(0, 1.9)$ y $X = mín(T, C)$.

2. $D2$: $T = Y \cdot \chi^2(3) + (1 - Y) \cdot \chi^2(60)$,
   $C = Y \cdot F(6, 3) + (1 - Y) \cdot weibull(4, 70)$ y $X = mín(T, C)$.

3. $D3$: $T = Y \cdot Exp(6) + (1 - Y) \cdot weibull(2, 3)$,
   $C = Y \cdot \chi^2(1) + (1 - Y) \cdot weibull(1, 2)$ y $X = mín(T, C)$.

donde $lognorm(\mu, \sigma^2)$ es la distribución *log-normal* con media μ y varianza σ^2 (del logaritmo natural de la variable que se distribuye lognormal).

Los correspondientes valores reales del coeficiente de Gini son: 0.4175, 0.5113 y 0.5938, respectivamente. Reportamos los resultados de la simulación en una tabla, igualmente error medio, error cuadrado medio y las probabilidades de cobertura.

Tabla 3.2: Resultados de simulación para censura dependiente de covariables

n	Estimador	Error	ECM(x10^2)	P.C.	Error	ECM(x10^2)	P.C.
		D1			D2		
300	Bon	-0.006	0.082	0.09	-0.224	5.09	0
	Ord.	0.037	0.175	0.296	0.031	0.152	0.446
	Ind.	0.007	0.094	0.922	-0.211	4.537	0
	Alt.Ind.	-0.006	0.082	0.66	-0.224	5.09	0
	Dep.	-0.020	0.121	0.348	-0.059	0.068	0.158
	Alt.Dep.	-0.006	0.080	0.656	-0.009	0.049	0.722
600	Bon	-0.003	0.043	0.076	-0.223	5.027	0
	Ord.	0.035	0.148	0.054	0.031	0.125	0.158
	Ind.	0.004	0.047	0.944	-0.217	4.744	0
	Alt.Ind.	-0.003	0.043	0.492	-0.223	5.027	0
	Dep.	-0.013	0.060	0.404	-0.036	0.023	0.166
	Alt.Dep.	-0.003	0.041	0.502	-0.004	0.032	0.586
1200	Bon	-0.004	0.020	0.032	-0.222	4.961	0
	Ord.	0.036	0.136	0	0.031	0.113	0.01
	Ind.	-0.0003	0.019	0.94	-0.219	4.816	0
	Alt.Ind.	-0.002	0.020	0.368	-0.222	4.961	0
	Dep.	-0.009	0.028	0.354	-0.021	0.018	0.212
	Alt.Dep.	-0.003	0.019	0.428	-0.002	0.015	0.448
		D3					
300	Bon	-0.043	0.248	0.012			
	Ord.	0.048	0.252	0.066			
	Ind.	-0.018	0.114	0.858			
	Alt.Ind.	-0.043	0.248	0.182			
	Dep.	-0.014	0.073	0.432			
	Alt.Dep.	-0.001	0.049	0.688			
600	Bon	-0.043	0.215	0.004			
	Ord.	0.047	0.233	0			
	Ind.	-0.030	0.128	0.592			
	Alt.Ind.	-0.043	0.215	0.028			
	Dep.	-0.012	0.040	0.376			
	Alt.Dep.	-0.003	0.026	0.496			
1200	Bon	-0.040	0.174	0			
	Ord.	0.047	0.226	0			
	Ind.	-0.033	0.127	0.29			

Tabla 3.2 continuación

Alt.Ind.	-0.040	0.174	0.002
Dep.	-0.008	0.018	0.408
Alt.Dep.	-0.001	0.012	0.384

Con los resultados de la tabla 3.2, para censura dependiente, podemos hacer algunas observaciones.

Este diseño del experimento es muy simple, por ello los estimadores para censura independiente pueden comportarse bien, pero es de destacar lo siguiente. Solo los estimadores para censura dependiente mejoran considerablemente con el incremento del tamaño de muestra, no se puede garantizar que bajo censura dependiente los estimadores para censura independiente sean también consistentes.

Cuando el valor verdadero del coeficiente de Gini es relativamente alto, todos los estimadores tienden a subestimarlo y por tanto los intervalos de confianza no logran captar el valor verdadero.

En este experimento como el anterior parece destacar el estimador $\widehat{G}_w^Z$, que en las tablas nombramos como **Alt.Dep.**. Sin embargo, no deja de ser destacable que esto suceda con el mismo ponderador $\frac{\delta_i}{\widehat{K}_C^{Z^i}(X_i)}$, analizado en este trabajo.

Nuevamente, al analizar las probabilidades de cobertura vemos que los intervalos de confianza no alcanzan a cubrir al valor verdadero en la proporción teórica. A pesar de la consistencia del estimador para dependencia y su distribución aproximadamente normal, parece mantenerse un sesgo que el artículo de Lv, G. Zhang y Ren 2017 no aborda.

Las gráficas ilustran mejor el punto anterior.

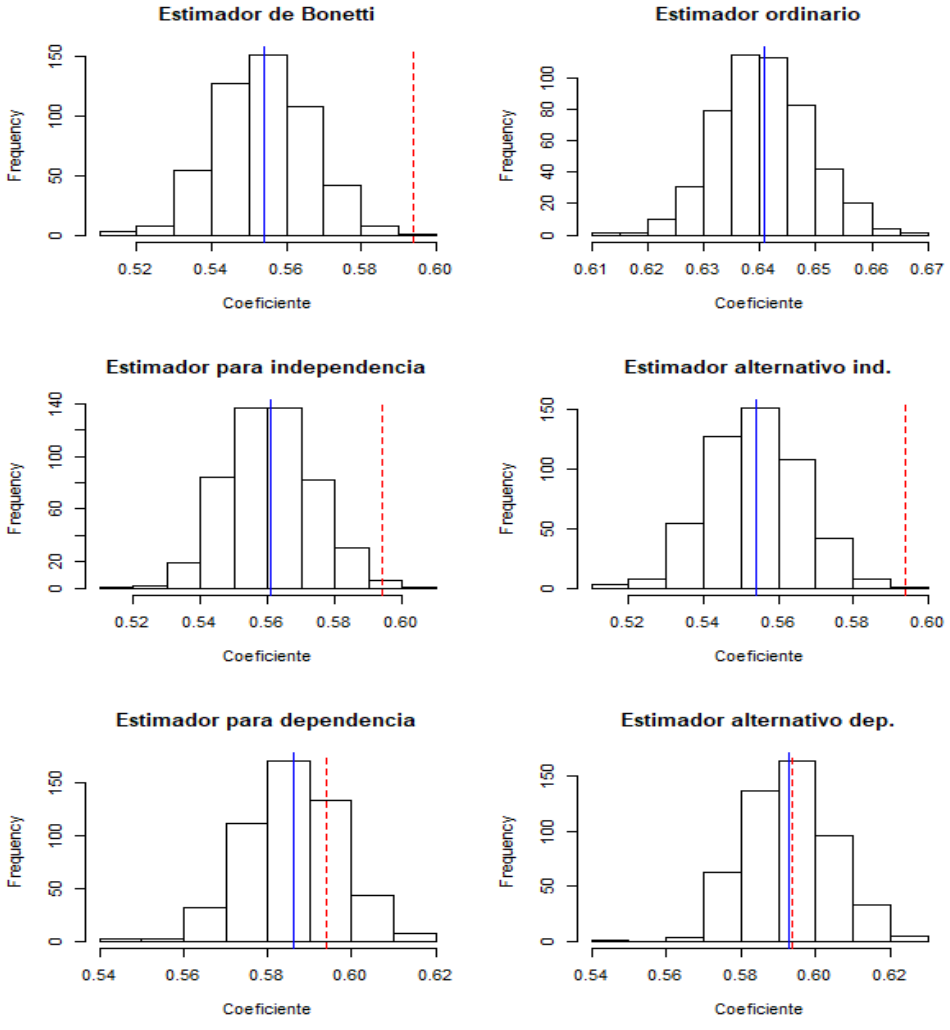

Figura 3.1: Histogramas de los estimadores del coeficiente de Gini para el tercer conjunto de datos simulados. La línea azul indica la media muestral de las estimaciones, mientras que la línea punteada roja indica el valor verdadero del coeficiente para este conjunto de datos.

La figura 3.1 ilustra como el estimador para dependencia es uno de los que mejor

aproximan el coeficiente verdadero, sin embargo es notorio que a pesar de su consistencia prevalece un sesgo en el estimador. Por este motivo y también debido a la rápida disminución de la varianza es que los intervalos de confianza fallan la mayoría de las ocasiones en cubrir al valor verdadero.

3.1.3. Análisis de Robustez

Con el análisis de robustez queremos mostrar qué tan afectada se ve la estimación del coeficiente de Gini para censura dependiente de covariables cuando se utilizan variables no adecuadas para el modelo o cuando el mismo modelo subyacente a la función de riesgo no es lineal (no se ajusta al modelo aditivo de Aalen).

El experimento se plantea de la siguiente manera. Primero generamos dos variables $Y_1 \sim \chi^2(6)$, $Y_2 \sim weibull(10, 13)$ y $T = Y1 + U(0, 1)^2$, donde $U(a, b)$ es la distribución uniforme con límites inferior a y, superior, b. El coeficiente de Gini correspondiente a T es aproximadamente 0.2973. Después establecemos 5 formas para la función de riesgo $h^i_C(x)$:

1. $h^i_C(x) = \beta_i Y_{1i}$, con $\beta_i \sim U(0, 0.04)$
2. $h^i_C(x) = \beta_i Y^2_{1i}$, con $\beta_i \sim weibull(0.6, 0.005)$
3. $h^i_C(x) = \beta_i Y_{1i} \cdot x/2$, con $\beta_i \sim U(0, 0.002)$
4. $h^i_C(x) = \beta_i Y^2_{1i} \cdot x/2$, con $\beta_i \sim U(0, 0.0003)$
5. $h^i_C(x) = 0.03e^{(\beta_i Y_{1i})}$, con $\beta_i \sim U(0, 0.1)$

C_i será tal que cumpla estas formas de censura, es decir cuando $C_i = x$ su función de riesgo, dada la covariable Y_{1i}, será alguna de las 5 enumeradas. C_i se puede generar de dos maneras, principalmente.

En los casos 1 y 2, las funciones de riesgo son constantes respecto del tiempo, esta forma se puede conseguir con la distribución exponencial cuya función de riesgo es igual a su parámetro de tasa, λ. Luego, si $X \sim Exp(\lambda)$, $kX \sim Exp(\lambda/k)$ para $k > 0$, entonces podemos generar $C_i = (1/\beta_i Y_{1i}) \cdot Exp(1)$, por ejemplo, para el caso número 1. Otra opción es generar una variable uniforme $u \sim U(0, 1)$ que coincide con la distribución de la función de supervivencia, si integramos la función de riesgo respecto del tiempo tenemos que es igual a $ln(S(x))$, usamos la función exponencial en ambos lados de esa igualdad y tenemos, por ejemplo para el caso 1 que $S(x) = e^{(-\beta_i Y_{1i} x)}$, que es la forma de la función de supervivencia de la distribución exponencial. C_i podemos generarla despejando x, es decir $C_i = -ln(u_i)/\beta_i Y_{1i}$.

En los casos 3 y 4, las funciones de riesgo no son constantes respecto al tiempo, pero viendo con cuidado podremos notar que siguen la forma de la función de riesgo de la distribución *weibull* con parámetro de forma $a = 2$ y parámetro de escala $b = 2$. De manera similar al caso de la distribución exponencial, si $X \sim weibull(a, b)$, entonces $kX \sim weibull(a, kb)$ para $k > 0$, por lo que, ejemplificando para el caso 1, $C_i = (1/\sqrt{\beta_i Y_{1i}}) \cdot weibull(2, 2)$. O también, por medio de una variable uniforme, $C_i = \sqrt{\frac{-4ln(u_i)}{\beta_i Y_{1i}}}$.

La forma de la función de riesgo del caso 5 corresponde a un modelo de riesgos proporcionales (modelo de Cox). Aquí la forma más simple de generar C_i es por medio de una variable uniforme, esto es $C_i = -ln(u_i)/(0.03e^{\beta_i Y_{1i}})$.

Una vez generadas las variables C correspondientes a cada caso, obtenemos como antes $X_i = min(T_i, C_i)$. De esta forma es clara la independencia condicional de T y C dada Y_1, mientras que las formas de la función de riego son, sólo en los casos 1 y 3, perteneciente al modelo aditivo de Aalen.

Para comprobar la robustez de los estimadores para dependencia utilizamos 3 combinaciones de $\boldsymbol{Z}_i$, que son: $\boldsymbol{Z}_i^T = (1, Y_{1i})$, $\boldsymbol{Z}_i^T = (1, Y_{1i}, Y_{2i})$ y $\boldsymbol{Z}_i^T = (1, Y_{2i})$. La primera combinación de covariables es la Adecuada (AD), la segunda Contiene variables Irrelevantes (CI) y la tercera se forma Sin variables Relevantes (SR).

Sólo en los casos 1 y 3 el modelo AD se ajusta perfectamente, en los casos 2 y 4 el modelo no es completamente preciso al no entrar Y_1 de manera lineal, lo que podría cambiar si $\boldsymbol{Z}_i^T = (1, Y_{1i}^2)$, pero se trata de probar la precisión de los estimadores bajo estas desviaciones por lo que no incluimos ese caso. En cuanto al caso 5, es evidente la diferencia con el modelo de Aalen.

Presentamos los mismos indicadores que en las primeras simulaciones, el error medio, el error cuadrado medio y las probabilidades de cobertura. Ahora tenemos porcentajes de censura de entre 15 y 30 por ciento, esto para que las muestras pudieran converger con la precisión buscada en tamaños de muestra de hasta $n = 500$. Por supuesto que dependiendo de las distribuciones de T y C, y el porcentaje de censura la convergencia puede requerir un tamaño de muestra más grande.

Tabla 3.3: Análisis de Robustez

n	Estimador	Error	ECM(x10^2)	P.C.	Error	ECM(x10^2)	P.C.
		D1			D2		

Tabla 3.3: Análisis de Robustez

n	Estimador	Error	ECM(x10^2)	P.C.	Error	ECM(x10^2)	P.C.
100	Bon	-0.019	0.105	0.104	-0.014	0.097	0.124
	Ord.	0.066	0.502	0.356	0.117	1.459	0.026
	Ind.	0.005	0.083	0.942	0.005	0.082	0.938
	Alt.Ind.	-0.019	0.105	0.93	-0.014	0.097	0.962
	Dep.AD	-0.017	0.400	0.562	-0.0004	0.360	0.528
	Alt.Dep.AD	-0.015	0.097	0.934	-0.011	0.118	0.934
	Dep.CI	-0.076	34.164	0.52	-0.010	1.129	0.532
	Alt.Dep.CI	-0.020	0.198	0.908	-0.013	0.125	0.942
	Dep.SR	-0.015	0.162	0.538	-0.003	0.096	0.51
	Alt.Dep.SR	-0.023	0.121	0.91	-0.016	0.101	0.958
500	Bon	-0.014	0.036	0.036	-0.010	0.024	0.026
	Ord.	0.068	0.471	0	0.119	1.442	0
	Ind.	-0.010	0.026	0.842	-0.006	0.064	0.892
	Alt.Ind.	-0.014	0.036	0.376	-0.010	0.024	0.544
	Dep.AD	-0.005	0.024	0.586	0.002	0.018	0.522
	Alt.Dep.AD	-0.004	0.022	0.570	-9e-05	0.027	0.562
	Dep.CI	-0.012	1.668	0.496	-0.001	0.070	0.506
	Alt.Dep.CI	0.018	28.095	0.546	-0.001	0.028	0.556
	Dep.SR	-0.014	0.040	0.288	-0.007	0.022	0.382
	Alt.Dep.SR	-0.016	0.041	0.35	-0.010	0.025	0.537
		D3			D4		
100	Bon	-0.009	0.049	0.134	-0.0180	0.073	0.084
	Ord.	-0.015	0.057	0.9	-0.025	0.095	0.76
	Ind.	0.003	0.042	0.95	-0.015	0.045	0.926
	Alt.Ind.	-0.009	0.049	0.918	-0.0180	0.073	0.836
	Dep.AD	-0.005	0.071	0.546	-0.015	0.120	0.476
	Alt.Dep.AD	-0.008	0.047	0.917	-0.014	0.060	0.873
	Dep.CI	-0.012	0.124	0.5	-0.026	0.199	0.398
	Alt.Dep.CI	-0.011	0.053	0.912	-0.016	0.073	0.862
	Dep.SR	-0.007	0.060	0.528	-0.017	0.089	0.398
	Alt.Dep.SR	-0.012	0.053	0.901	-0.021	0.082	0.819
500	Bon	-0.007	0.013	0.046	-0.014	0.029	0.014
	Ord.	-0.014	0.026	0.272	-0.023	0.058	0.052
	Ind.	-0.005	0.013	0.906	-0.012	0.025	0.73
	Alt.Ind.	-0.007	0.013	0.516	-0.014	0.029	0.226
	Dep.AD	-0.004	0.010	0.512	-0.008	0.022	0.436
	Alt.Dep.AD	-0.003	0.009	0.572	-0.006	0.013	0.524
	Dep.CI	-0.026	24.656	0.476	-0.013	0.053	0.356
	Alt.Dep.CI	-0.005	0.051	0.552	-0.008	0.022	0.468
	Dep.SR	0.007	0.015	0.416	-0.015	0.035	0.158

Tabla 3.3: Análisis de Robustez

n	Estimador	Error	ECM(x10^2)	P.C.	Error	ECM(x10^2)	P.C.
	Alt.Dep.SR	0.008	0.014	0.482	-0.016	0.033	0.185
		D5					
100	Bon	-0.012	0.062	0.128			
	Ord.	0.026	0.113	0.864			
	Ind.	0.003	0.051	0.946			
	Alt.Ind.	-0.012	0.062	0.942			
	Dep.AD	-0.004	0.081	0.57			
	Alt.Dep.AD	-0.011	0.056	0.954			
	Dep.CI	-0.014	0.259	0.528			
	Alt.Dep.CI	-0.018	1.784	0.944			
	Dep.SR	-0.005	0.068	0.508			
	Alt.Dep.SR	-0.014	0.067	0.936			
500	Bon	-0.007	0.015	0.128			
	Ord.	0.029	0.097	0.864			
	Ind.	-0.004	0.014	0.946			
	Alt.Ind.	-0.007	0.015	0.942			
	Dep.AD	-0.003	0.012	0.57			
	Alt.Dep.AD	-0.004	0.011	0.954			
	Dep.CI	-0.006	0.022	0.528			
	Alt.Dep.CI	-0.005	0.013	0.944			
	Dep.SR	-0.007	0.016	0.508			
	Alt.Dep.SR	0.017	-0.008	0.936			

A partir de la tabla 3.3 podemos hacer las siguientes observaciones. Para las 5 formas de la función de riesgo los estimadores de dependencia del modelo AD fueron los más precisos según el error medio y el error cuadrado medio. Al tener poca proporción de censura, los demás estimadores también se muestran cercanos al valor real, pero sin ser tan precisos como los estimadores para dependencia del modelo adecuado, incluso cuando la forma de la función de riesgo subyacente es totalmente diferente al modelo aditivo.

El modelo que incluye una variable irrelevante (CI) se comportó bien en lo general para los 5 casos de acuerdo con su error medio que indica que en promedio no se aleja mucho del valor real, sin embargo, de acuerdo al error cuadrado medio podemos decir que para estimaciones individuales puede representar sesgos importantes.

En cuanto a las estimaciones que no incluyen variables relevantes, destacadamente se comportan mejor que en el caso de que se mezcle información relevante con irrele-

vante. Esto a bajos niveles de censura.

3.2. Desigualdad en el tiempo para obtener empleo entre mexicanos con algún posgrado

Medir la desigualdad en el tiempo para encontrar trabajo es una herramienta útil para detectar un problema. Imaginemos que solo una proporción pequeña de personas que buscan empleo lo encuentren en pocas semanas, mientras que la mayoría tardan meses, incluso años para encontrar trabajo.

Un análisis como el que propone este trabajo es parte del primer paso: el diagnóstico, ¿existe un problema de desigualdad en esta tarea de colocarse en un empleo? La muestra nos dirá inmediatamente que tan cortos o largos son los periodos que pasa una persona buscando trabajo. Pero una desigualdad marcada, advertiría que puede haber un problema estructural además de que podríamos detectar algunos de los factores que se relacionan con esas diferencias.

Antes de iniciar con el análisis de nuestra aplicación, debemos dejar en claro que es un primer análisis del problema que sin duda puede mejorarse para obtener una medición más fidedigna y precisa, pero no por ello sin valor. Este análisis inicia para nuestro país el abordaje del problema del desempleo desde un enfoque que no se había aplicado, que arroja una medida fácil de interpretar y que resume varias temáticas alrededor del problema del desempleo pues nos habla de la distribución y no es solo un resultado agregado, como el tiempo medio de desempleo.

También debemos hacer algunos señalamientos sobre cómo interpretar el resultado. Recordemos que el coeficiente de Gini se basa en la curva de Lorenz, esta curva presenta de manera proporcional la distribución de la variable de interés en el eje vertical y la población correspondiente en el eje horizontal. La población está ordenada de menor a mayor según su medida en la variables de interés. Cuando la variable de interés es el tiempo, por ejemplo, los primeros cuantiles tienen un tiempo registrado menor que los últimos cuantiles.

Ya que un tiempo corto para conseguir empleo se considera como positivo, dado un tiempo considerado como largo (cuestión aparte, que corresponde teorizar a la Economía), un grado alto de desigualdad significa que pocas personas “privilegiadas” registran tiempos cortos para conseguir empleo, mientras que la mayoría registran tiempos largos.

Por lo anterior, cuando la variable de interés es el tiempo, un coeficiente de Gini bajo nos diría que hay una desigualdad preocupante o que debe tomarse como negativa, mientras que un coeficiente de Gini alto sería algo positivo, hablando en términos generales. Para entender mejor este punto vamos a ilustrarlo con un ejemplo:

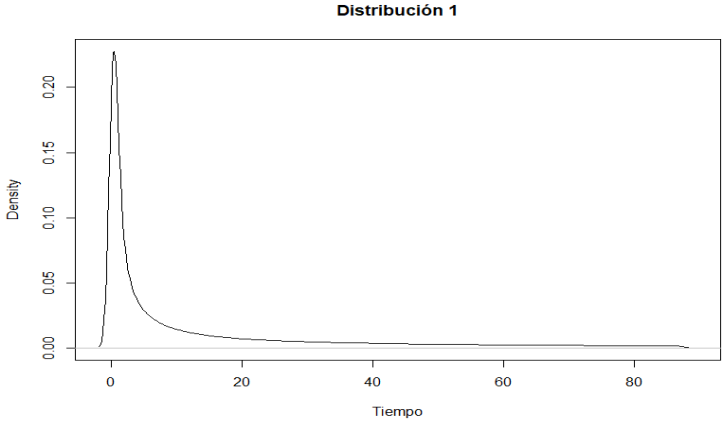

Figura 3.2: Gráfica de densidad de una distribución ejemplo.

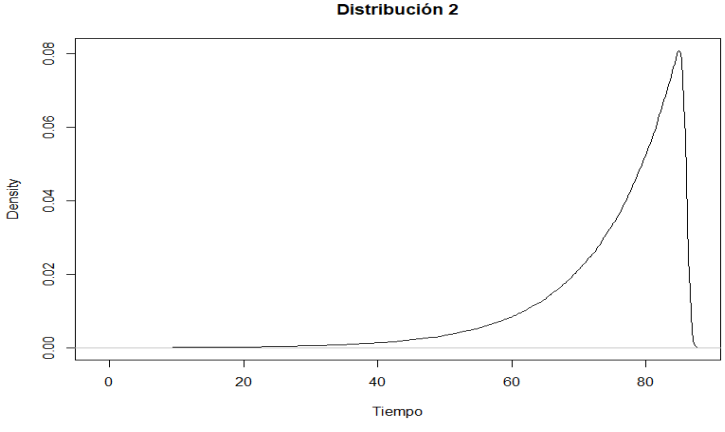

Figura 3.3: Gráfica de densidad de una distribución ejemplo.

Las gráficas anteriores nos muestran dos distribuciones contrarias de la variable

tiempo. En el primer caso, la mayoría de los individuos tienen tiempos cortos y son pocos los que tienen tiempos largos, lo que sería algo positivo hablando de tiempos para conseguir empleo. En la segunda gráfica vemos lo opuesto, pocos individuos registran tiempos cortos y la mayoría registran tiempos largos, lo que claramente se consideraría negativo. Pero veamos ahora que los coeficientes de Gini son 0.7155 y 0.0728, respectivamente.

Ya han quedado expresadas estas consideraciones, podemos pasar ahora a la presentación de los datos empíricos. El ejemplo que vamos a presentar en este trabajo es la **desigualdad en el tiempo que tardan los egresados de algún posgrado para encontrar empleo**.

Por medio de un cuestionario formulado por el autor de este trabajo se obtuvo una muestra de 107 sujetos. Registramos el tiempo transcurrido hasta conseguir empleo o una actividad de tiempo completo, principalmente ingresar a un nuevo posgrado. Como covariables registramos el sexo, el promedio obtenido en el posgrado, sueldo mínimo mensual deseado y si se es egresado de escuela privada o pública.

El sondeo fue realizado por medio de un grupo de "Facebook" y esto tiene una explicación. El objetivo claro del sondeo fue medir la desigualdad en el tiempo que tarda un mexicano con posgrado en conseguir empleo. El medio fue un grupo cerrado en Facebook al que se unen principalmente personas que estudian o estudiaron un posgrado con el apoyo de una beca del CONACYT, llamado "*Becarios CONACYT*". Si bien una consulta a través de internet no garantiza tener una muestra representativa de la población objetivo, las características de este grupo nos dicen que hay buenas posibilidades de obtener observaciones representativas. El grupo tiene una cantidad importante de miembros (cerca de 80 mil), de todas partes de la república e incluso usuarios que radican en otros países. No es irreal asumir que la mayoría de los egresados de las últimas generaciones tienen acceso a internet y utilizan este medio para comunicarse, informarse e interactuar. La población objetivo no es tan grande pues no buscamos analizar a la fuerza de trabajo de todo el país sino solo a los mexicanos con posgrado, lo que nos hace pensar que la muestra obtenida no debe ser tan sesgada.

La muestra tiene un 44 % de observaciones censuradas. El 64.5 % de quienes respondieron la encuesta son mujeres. Preguntamos el tiempo en semanas desde el término de un posgrado hasta conseguir un trabajo o iniciar una actividad diferente de tiempo completo, como por ejemplo: comenzar un nuevo posgrado, abrir un negocio, etc. El tiempo promedio para colocarse en una nueva actividad de tiempo completo entre los encuestados fue de 22.4 semanas. Aquí, T es el tiempo para conseguir un empleo, mientras que C es el tiempo para colocarse en una actividad distinta de tiempo completo, destacadamente empezar un nuevo posgrado.

Solicitamos el tiempo en semanas porque pensamos que una medida más corta, aunque más precisa, sería desalentadora para algunos participantes, mientras que las semanas son más rápidas para contar. Por otro lado, los estimadores requieren continuidad de la variable X, así que sustituimos el valor contestado por una variable uniforme en el rango de la respuesta original ± 0.5. Este manejo no altera la distribución de los datos o, mejor dicho, el cambio es insignificante, lo que se puede ver en que el promedio es prácticamente el mismo y el coeficiente de Gini ordinario es casi igual por milésimas.

A continuación, mostramos los histogramas de las distribuciones

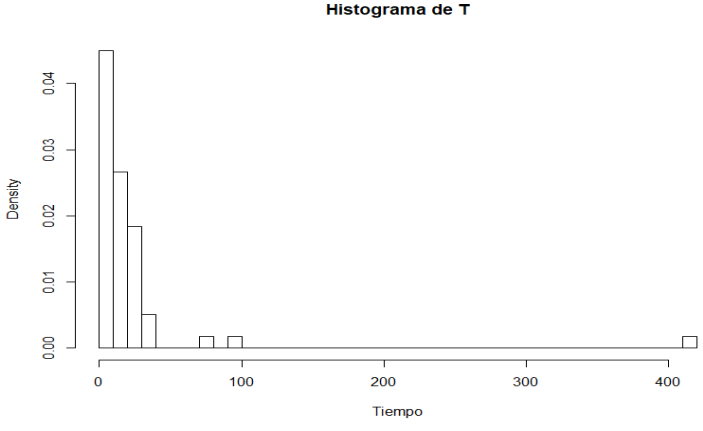

Figura 3.4: Histograma del tiempo para conseguir empleo.

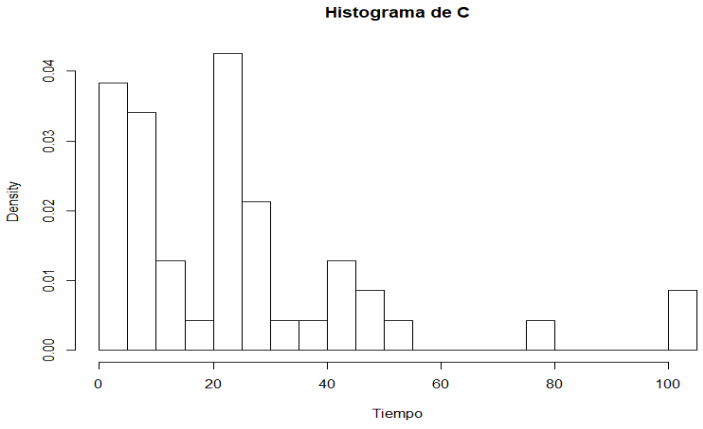

Figura 3.5: Histograma de los tiempo de censura.

De las covariables registradas, las que mejor explican la censura son el sexo y con menos explicación si se es egresado de universidad Privada o Pública. Para notar esto nos apoyamos del paquete de `R`, `timereg`(Scheike y M.-J. Zhang 2011), con el que podemos ajustar un modelo aditivo de Aalen a nuestros datos. Utilizando todas las variables el resumen del ajuste arroja lo siguiente

```
> summary(fit)
Additive Aalen Model

Test for nonparametric terms

Test for non-significant effects
             Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                                0                   0
M$Sexo                                     0                   0
M$Universidad                              0                   0
M$Promedio                                 0                   0
M$Sueldo                                   0                   0
```

graficando las funciones de regresión acumuladas o coeficientes acumulados de regresión, $B(x)$ nos damos cuenta de que la covariable Sueldo, que es el sueldo mínimo deseado para aceptar un empleo, no tiene efectos sobre el tiempo de censura, así que la retiramos del modelo y resulta ahora

```
> summary(fit)
Additive Aalen Model

Test for nonparametric terms

Test for non-significant effects
             Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                            Inf                   0
M$Sexo                                 Inf                   0
M$Universidad                          Inf                   0
M$Promedio                             Inf                   0
```

parece que estas tres variables son significativas, lo indica el "Supremum-test". Para mayor profundidad en el uso de este paquete y esta función pude consultar Martinussen y H. Scheike 2006.

Evaluándolas por separado nos damos cuenta de que la influencia de la covariable "Promedio" (significa el promedio obtenido en el posgrado) no es claramente determinante al explicar el tiempo de censura y por ello la retiramos

```
> summary(fit3)
Additive Aalen Model

Test for nonparametric terms

Test for non-significant effects
            Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                          2.29               0.223
M$Promedio                           2.56               0.128
```

Con el par de covariables "Sexo" y "Universidad" realizamos las estimaciones del coeficiente de Gin y obtenemos los siguientes resultados.

Tabla 3.4: Coeficiente de Gini para duración del desempleo entre mexicanos egresados de un posgrado

Estimador	Valor	Sexo	Universidad	I.C.(95 %)
Bon.	0.6997			(0.694,0.705)
Ord.	0.5994			(0.481,0.717)
Ind.	0.8530			(0.753,0.952)
Alt. Ind.	0.6997			(0.584,0.815)
Dep.	0.8468	✓		(0.350,0.889)
Alt.Dep.	0.6988	✓		(0.589,0.808)
Dep.	0.8244		✓	(0.364,0.894)
Alt.Dep.	0.7052		✓	(0.595,0.815)
Dep.	0.7857	✓	✓	(0.367,0.896)
Alt.Dep.	0.7090	✓	✓	(0.597,0.820)

Por los resultados de tabla 3.4 y en congruencia con los resultados de las simulaciones podemos decir que es muy probable que el coeficiente de Gini para la duración del desempleo entre egresados de un posgrado se encuentre un poco arriba de 0.71. Este coeficiente indica que la mayor parte de las observaciones registran tiempos relativamente bajos, lo que puede considerarse positivo.

3.3. Conclusiones

Sin duda el experimento debería repetirse con diferentes muestras y de tamaño mayor para conocer de manera más precisa este estimador. Es también importante incluir o acompañar este análisis con un estudio específico sobre las covariables que afectan tanto el tiempo para conseguir empleo como el tiempo de censura cuando se ingresa en otra actividad, muchas de estas covariables seguramente no las incluimos en la encuesta.

Los estimadores de Lv, G. Zhang y Ren 2017 demuestran comportarse bien en lo general al ser bastante precisos y tener propiedades de robustez. Lo más destacado es su parte de ponderación de probabilidad inversa que ajusta bien a otros estimadores como el estimador (3.1) que brinda también muy buenos resultados.

No obstante en este trabajo pudimos corroborar por medio de simulación que el estimador para dependencia de covariables parece mantener un sesgo, a pesar de la consistencia del estimador y del aumento del tamaño de la muestra, como vimos en las gráficas de la figura 3.1. Este es un problema que el trabajo de Lv, G. Zhang y Ren 2017 no aborda, más aun parece no existir, pero al probar con diferentes conjuntos

de datos surge de inmediato.

El motivo del sesgo y su correspondiente corrección va más allá de los límites de este trabajo, pero queda como un tema abierto para posteriores trabajos alrededor del coeficiente de Gini bajo censura dependiente de covariables.

Apéndice A

Código en R

A.1. Funciones

```
#####################  FUNCIONES   ######################
#Declaramos funciones hechas por el usuario para hacer mas fluido su uso
#y más limpio el programa principal
############## Funión de distribución acumulada empírica #############
Fn<-function(x,M,k){mean((M$delta/k)*((M$X<=x)*1),na.rm=TRUE)}
######### Estimación de riesgo acumulada #######
Hsi<-function(M,z=z,n=n){
c1<-length(z[1,]); Ri<-sapply(M$X,FUN=function(x) (M$X>=x)*1)
N<-sapply(M$X,FUN=function(x) (M$X<=x&M$delta==0)*1)
dB<-matrix(0,c1,n)
for(i in 1:n){
  A<-(t(Ri[,i]*z))%*%(Ri[,i]*z)
  if(qr(A)$rank==length(z[1,])){
    Ainv<-solve(A)
  }else{
    P<-eigen(A)$vectors; E<-eigen(A)$values
    Epls<-matrix(0,c1,c1)
    for(j in 1:c1){if(E[j]==0){Epls[j,j]<-0}else{Epls[j,j]<-1/E[j]}}
    Ainv<-P%*%Epls%*%t(P)
  }
  dB[,i]<-Ainv%*%z[i,]
}
dH<-t(N)*(z%*%dB);Hci<-c()
for(i in 1:n){
  Hci[i]<-sum(dH[i,1:i])
}
return(Hci)
}
######### Estimación de la función de supervivencia condicinal ######
kmz<-function(M,z=z,n=n){
c1<-length(z[1,]); Ri<-sapply(M$X,FUN=function(x) (M$X>=x)*1)
N<-sapply(M$X,FUN=function(x) (M$X<=x&M$delta==0)*1)
dB<-matrix(0,c1,n); kz<-c()
for(i in 1:n){
```

```
  A<-(t(Ri[,i]*z))%*%(Ri[,i]*z)
  if(qr(A)$rank==length(z[1,])){
    Ainv<-solve(A)
  }else{
    P<-eigen(A)$vectors; E<-eigen(A)$values
    Epls<-matrix(0,c1,c1)
    for(j in 1:c1){if(E[j]==0){Epls[j,j]<-0}else{Epls[j,j]<-1/E[j]}}
    Ainv<-P%*%Epls%*%t(P)
  }
  dB[,i]<-Ainv%*%z[i,]
}
dH<-t(N)*(z%*%dB)
for (j in 1:n) {
  kz[j]<-prod(1-dH[j,1:j])
}
return(kz)
}
######### Estimador del coef de Gini de Bonnetti ###########
GBonn<-function(data,Tmax=max(data[,1])){
  data1 <- data.frame(data)
  info <- survfit(Surv(data[,1], data[,2])~1 , type='kaplan-meier', data=data1)
  K<-length(info$time)
  S<-matrix(info$surv, ncol=1, nrow=K)
  num<-matrix(ncol=1,nrow=K)
  T<-matrix(info$time,ncol=1, nrow=K)
  num[1]<-1*T[1]
  if (K>=2){
    for (i in 2:K)
      num[i]<-num[i-1]+((S[i-1])^2)*(T[i]-T[i-1])
  }
  den <- matrix(ncol=1,nrow=K)
  den[1] <- 1*T[1]
  if (K>=2){
    for (i in 2:K)
      den[i]<-den[i-1]+S[i-1]*(T[i]-T[i-1])
  }
  G <- 1-(num/den)
  indices <- sum(1*(T<Tmax))
  lastpiecesnum <- ((S[indices])^2)*(Tmax-T[indices])
  lastpiecesden <- (S[indices])*(Tmax-T[indices])
  GTmax <- 1-(num[indices]+lastpiecesnum)/(den[indices]+lastpiecesden)
  return(list( G=G,info=info,GTmax=GTmax))
}
########## Varianza del estimador de Bonnetti ###########
VGBonn<-function(data, Tmax=max(data[,1])){
  data1<-data.frame(data)
  info<- survfit(Surv(data1[,1], data1[,2])~1 , type='kaplan-meier', data=data1)
  S<-matrix(info$surv)
  T<-matrix(info$time)
  indices <- sum(1*(T<=Tmax))
  if (indices==0){
    var<-0
    return(var=var)
  }
  else{
    Vt<-matrix(ncol=1,nrow=indices)
    Vt[1]<-1*T[1]
    if(indices>=2){
      for (i in 2:indices){
        Vt[i]<-Vt[i-1]+((S[i-1])^2)*(T[i]-T[i-1])
```

```
      }
    }
    lastpiecesVt<-((S[indices])^2)*(Tmax-T[indices])
    VtMax<-Vt[indices]+lastpiecesVt
    Wt<-matrix(ncol=1,nrow=indices)
    Wt[1]<-1*T[1]
    if(indices>=2){
      for (i in 2:indices){
        Wt[i]<-Wt[i-1]+S[i-1]*(T[i]-T[i-1])
      }
    }
    lastpiecesWt<-(S[indices])*(Tmax-T[indices])
    WtMax<-Wt[indices]+lastpiecesWt
    mu2<-matrix(ncol=1,nrow=indices)
    mu2<-VtMax-Vt
    mu<-matrix(ncol=1,nrow=indices)
    mu<-WtMax-Wt
    n<-length(data[,1])
    event<-matrix( ncol=1, nrow=indices)
    atrisk<-matrix(ncol=1, nrow=indices)
    for (i in 1:indices){
      event[i]<-info$n.event[i]
      atrisk[i]<-info$n.risk[i]
    }
    dsigma<-matrix(0, ncol=1,nrow=indices)
    for (i in 1:indices){
      if (atrisk[i]>0)
        dsigma[i]<-(n*event[i])/(atrisk[i]^2)
    }
    varistant<-matrix(ncol=1, nrow=indices)
    for (i in 1:indices){
      varistant[i]<-(4*exp(2*log(mu2[i])-2*log(WtMax))+exp(2*log(mu[i])
                    +2*log(VtMax)-4*log(WtMax))-4*exp(log(mu[i])+log(mu2[i])
                     +log(VtMax)-3*log(WtMax)))*(dsigma[i])
    }
    var<-matrix(ncol=1,nrow=indices)
    var[1]<-varistant[1]
    if(indices>=2){
      for (i in 2:indices){
        var[i]<-var[i-1]+varistant[i]
      }
    }
    return(var[indices]/n)
  }
}

##### Varianza de nuestro estimador  ##########
VGest<-function(M,km_c,km_t,n,Fcn){
  p<-M$delta/km_c; p02<-M$delta_0/km_c^2; prom<-1/(n*km_t)
  Ind<-sapply(M$X,FUN = function(x) (1*(M$X>=x)))
  Fx<-prom*apply((p*M$X*Ind),2,sum)
  Fx2<-prom*apply((p*(M$X^2)*Ind),2,sum)
  Xix<-M$X*Fcn+apply((p*M$X*Ind),2,mean)
  FXi<-prom*apply((p*Xix*Ind),2,sum)
  FXi2<-prom*apply((p*(Xix^2)*Ind),2,sum)
  FxFx<-prom*Fx*apply((p*M$X*Ind),2,sum)
  FxXi<-prom*apply((p*M$X*Xix*Ind),2,sum)
  FxFXi<-prom*FXi*apply((p*M$X*Ind),2,sum)
  FXiFx<-prom*Fx*apply((p*Xix*Ind),2,sum)
  FXiFXi<-prom*FXi*apply((p*Xix*Ind),2,sum)
```

```
  mu<-mean(p*M$X,na.rm=TRUE)
  xi<-mean(p*(2*Fcn*M$X),na.rm=TRUE)
  G<-(xi/mu)-1
  v1<-(1/(n-1))*sum(p*(2*(Xix-mean(p*Xix,na.rm=TRUE))
                       -(G+1)*(M$X-mu))^2,na.rm=TRUE)

  var<-(1/mu^2)*(v1 + 4*(mean(p02*FXi^2,na.rm=TRUE)+
      mean(p02*FXi2,na.rm=TRUE)-2*mean(p02*FXiFXi,na.rm=TRUE))+
      ((G+1)^2)*(mean(p02*Fx^2,na.rm=TRUE)+mean(p02*Fx2,na.rm=TRUE)-
                   2*mean(p02*FxFx,na.rm=TRUE))-
      4*(G+1)*(mean(p02*Fx*FXi,na.rm=TRUE)+mean(p02*FxXi,na.rm=TRUE)-
                 mean(p02*FXiFx,na.rm=TRUE)-mean(p02*FxFXi,na.rm=TRUE)))
  return(var)
}
####################################
```

A.2. Simulación

```
######################     SIMULACIONES      ##########################
library(survival)
library(Survgini)
library(laeken)

############ Corremos funciones construídas por usuario ##############
##### SIMULACIÓN PARA CENSURA INDEPENDIENTE #########
#########################################################################

#### 1. Primer conjunto de datos para censura independiente ####
#Generación de Población, Valor real y Muestra
D1.ind<-data.frame(T=rchisq(1E7,20),C=rexp(1E7,0.035),Y=rweibull(1E7,2,20))
plot(density(D1.ind$T)); lines(density(D1.ind$C)); lines(density(D1.ind$Y))
G1i.real<-(gini(D1.ind$T)$value)/100#Con paquete "laeken"

##### Ejemplo 1.1.- REPETICIÓN 500 veces con muestra tamaño 100 #####
propC11i<-c(); G11i.Bon<-c(); G11iIn.Bonn<-c(); G11i.esti<-c(); G11iIn.esti<-c()
G11i.ord<-c(); G11iIn.ord<-c(); G11i.pond<-c(); G11iIn.pond<-c(); G11i.dep<-c()
G11iIn.dep<-c(); G11i.pdep<-c(); G11iIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1i<-D1.ind[sample(nrow(D1.ind),1E2),]
M1i$X<-apply(cbind(M1i$T,M1i$C),1,min)
M1i$delta<-(M1i$X %in% D1.ind$T)*1
M1i$delta_0<-(M1i$X %in% D1.ind$C)*1
M1i<-M1i[order(M1i$X),]
km_c<-survfit(Surv(M1i$X,M1i$delta_0)~1,type='kaplan-meier',
              data=M1i)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5){
  M1i[length(km_c),1:3]<-D1.ind[sample(nrow(D1.ind),1),]
  M1i$X[n]<-min(M1i$T[n],M1i$C[n])
  M1i$delta[n]<-(M1i$X[n] %in% D1.ind$T)*1
  M1i$delta_0[n]<-(M1i$X[n] %in% D1.ind$C)*1
  M1i<-M1i[order(M1i$X),]
  km_c<-survfit(Surv(M1i$X,M1i$delta_0)~1,
                type='kaplan-meier',data=M1i)$surv}
```

```
  propC11i[k]<-mean((M1i$X %in% D1.ind$C)*1)#Prop. de obs. censuradas
######Estimadores
  z<-cbind(rep(1,n),M1i$Y)
  km_t<-survfit(Surv(M1i$X,M1i$delta)~1,type='kaplan-meier',
                data=M1i)$surv
  Fcn<-sapply(M1i$X,FUN=Fn,M=M1i,k=km_c)
  Bs<-Hs(M1i,z,n); km<-exp(-(z%*%Bs)); kzc<-apply(km,2,mean,na.rm=TRUE)
  kzc<-kmz(M1i,z,n)
  pesos<-M1i$delta/km_c; pesos.z<-M1i$delta/kzc
  Fzcn<-sapply(M1i$X,FUN=Fn,M=M1i,k=kzc)
  xi<-mean(pesos*(2*Fcn*M1i$X),na.rm=TRUE)
  mu<-mean(pesos*M1i$X,na.rm=TRUE)
  xiz<-mean(pesos.z*(2*Fzcn*M1i$X),na.rm=TRUE)
  muz<-mean(pesos.z*M1i$X,na.rm=TRUE)
#############    RESULTADOS  #################
G11i.Bon[k]<-GBonn(data = cbind(M1i$X,M1i$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1i$X,M1i$delta))
IC.GBon<-c(G11i.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G11i.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G11iIn.Bonn[k]<-(G1i.real>=IC.GBon[1]&G1i.real<=IC.GBon[2])*1

G11i.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1i,km_c,km_t,n,Fcn)
IC.Gesti<-c(G11i.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G11i.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G11iIn.esti[k]<-(G1i.real>=IC.Gesti[1]&G1i.real<=IC.Gesti[2])*1


ginord<-gini(M1i$X)
ginord<-variance(M1i$X,indicator=ginord,R=200,bootType="naive")
G11i.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G11i.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G11i.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G11iIn.ord[k]<-(G1i.real>=IC.Gord[1]&G1i.real<=IC.Gord[2])*1


ginpond<-gini(M1i$X,weights=pesos)
ginpond<-variance(M1i$X,indicator=ginpond,R=200,bootType="naive")
G11i.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G11i.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G11i.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G11iIn.pond[k]<-(G1i.real>=IC.Gpond[1]&G1i.real<=IC.Gpond[2])*1

G11i.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1i[sample(nrow(M1i),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y)
  Bs<-Hs(Muest,zb,nb); km<-exp(-(zb%*%Bs)); kzc<-apply(km,2,mean,na.rm=TRUE)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G11iIn.dep[k]<-(G1i.real>=IC.Gdep[1]&G1i.real<=IC.Gdep[2])*1

ginpdep<-gini(M1i$X,weights=pesos.z)
```

```
ginpdep<-variance(M1i$X,indicator=ginpdep,R=200,bootType="naive")
G11i.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G11i.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G11i.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G11iIn.pdep[k]<-(G1i.real>=IC.Gpdep[1]&G1i.real<=IC.Gpdep[2])*1
}
proc.time()-t


#####################################################################

#### 2. Segundo conjunto de datos para censura independiente ####
#Generación de Población, Valor real y Muestra
D2.ind<-data.frame(T=rexp(1E7,0.3),C=rf(1E7,20,2),Y=rweibull(1E7,2,20))
plot(density(D2.ind$T)); lines(density(D2.ind$C)); lines(density(D2.ind$Y))
G2i.real<-(gini(D2.ind$T)$value)/100#Con paquete "laeken"

##### Ejemplo 2.1.- REPETICIÓN 500 veces con muestra tamaño 100 #####
propC21i<-c(); G21i.Bon<-c(); G21iIn.Bonn<-c(); G21i.esti<-c(); G21iIn.esti<-c()
G21i.ord<-c(); G21iIn.ord<-c(); G21i.pond<-c(); G21iIn.pond<-c(); G21i.dep<-c()
G21iIn.dep<-c(); G21i.pdep<-c(); G21iIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1i<-D2.ind[sample(nrow(D2.ind),1E2),]
M1i$X<-apply(cbind(M1i$T,M1i$C),1,min)
M1i$delta<-(M1i$X %in% D2.ind$T)*1
M1i$delta_0<-(M1i$X %in% D2.ind$C)*1
M1i<-M1i[order(M1i$X),]
km_c<-survfit(Surv(M1i$X,M1i$delta_0)~1,type='kaplan-meier',
              data=M1i)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5){
  M1i[length(km_c),1:3]<-D2.ind[sample(nrow(D2.ind),1),]
  M1i$X[n]<-min(M1i$T[n],M1i$C[n])
  M1i$delta[n]<-(M1i$X[n] %in% D2.ind$T)*1
  M1i$delta_0[n]<-(M1i$X[n] %in% D2.ind$C)*1
  M1i<-M1i[order(M1i$X),]
  km_c<-survfit(Surv(M1i$X,M1i$delta_0)~1,
                type='kaplan-meier',data=M1i)$surv}

propC21i[k]<-mean((M1i$X %in% D2.ind$C)*1)#Prop. de obs. censuradas
######Estimadores
z<-cbind(rep(1,n),M1i$Y)
km_t<-survfit(Surv(M1i$X,M1i$delta)~1,type='kaplan-meier',
              data=M1i)$surv
Fcn<-sapply(M1i$X,FUN=Fn,M=M1i,k=km_c)
Bs<-Hs(M1i,z,n); km<-exp(-(z%*%Bs)); kzc<-apply(km,2,mean,na.rm=TRUE)
kzc<-kmz(M1i,z,n)
pesos<-M1i$delta/km_c; pesos.z<-M1i$delta/kzc
Fzcn<-sapply(M1i$X,FUN=Fn,M=M1i,k=kzc)
xi<-mean(pesos*(2*Fcn*M1i$X),na.rm=TRUE)
mu<-mean(pesos*M1i$X,na.rm=TRUE)
xiz<-mean(pesos.z*(2*Fzcn*M1i$X),na.rm=TRUE)
muz<-mean(pesos.z*M1i$X,na.rm=TRUE)
#############    RESULTADOS  #################
G21i.Bon[k]<-GBonn(data = cbind(M1i$X,M1i$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1i$X,M1i$delta))
IC.GBon<-c(G21i.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G21i.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
```

```
G21iIn.Bonn[k]<-(G2i.real>=IC.GBon[1]&G2i.real<=IC.GBon[2])*1

G21i.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1i,km_c,km_t,n,Fcn)
IC.Gesti<-c(G21i.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G21i.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G21iIn.esti[k]<-(G2i.real>=IC.Gesti[1]&G2i.real<=IC.Gesti[2])*1


ginord<-gini(M1i$X)
ginord<-variance(M1i$X,indicator=ginord,R=200,bootType="naive")
G21i.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G21i.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G21i.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G21iIn.ord[k]<-(G2i.real>=IC.Gord[1]&G2i.real<=IC.Gord[2])*1


ginpond<-gini(M1i$X,weights=pesos)
ginpond<-variance(M1i$X,indicator=ginpond,R=200,bootType="naive")
G21i.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G21i.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G21i.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G21iIn.pond[k]<-(G2i.real>=IC.Gpond[1]&G2i.real<=IC.Gpond[2])*1

G21i.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1i[sample(nrow(M1i),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y)
  Bs<-Hs(Muest,zb,nb); km<-exp(-(zb%*%Bs)); kzc<-apply(km,2,mean,na.rm=TRUE)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G21iIn.dep[k]<-(G2i.real>=IC.Gdep[1]&G2i.real<=IC.Gdep[2])*1

ginpdep<-gini(M1i$X,weights=pesos.z)
ginpdep<-variance(M1i$X,indicator=ginpdep,R=200,bootType="naive")
G21i.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G21i.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G21i.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G21iIn.pdep[k]<-(G2i.real>=IC.Gpdep[1]&G2i.real<=IC.Gpdep[2])*1
}
proc.time()-t


#####################################################################
#### 3. Tercer conjunto de datos para censura independiente ####
#Generación de Población, Valor real y Muestra
D3.ind<-data.frame(T=rf(1E7,3,3),C=rexp(1E7,1),Y=rweibull(1E7,2,20))
plot(density(D3.ind$T),xlim=c(0,200)); lines(density(D3.ind$C))
G3i.real<-(gini(D3.ind$T)$value)/100#Con paquete "laeken"

##### Ejemplo 3.1.- REPETICIÓN 500 veces con muestra tamaño 100 #####
propC31i<-c(); G31i.Bon<-c(); G31iIn.Bonn<-c(); G31i.esti<-c(); G31iIn.esti<-c()
```

```
G31i.ord<-c(); G31iIn.ord<-c(); G31i.pond<-c(); G31iIn.pond<-c(); G31i.dep<-c()
G31iIn.dep<-c(); G31i.pdep<-c(); G31iIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1i<-D3.ind[sample(nrow(D3.ind),1E2),]
M1i$X<-apply(cbind(M1i$T,M1i$C),1,min)
M1i$delta<-(M1i$X %in% D3.ind$T)*1
M1i$delta_0<-(M1i$X %in% D3.ind$C)*1
M1i<-M1i[order(M1i$X),]
km_c<-survfit(Surv(M1i$X,M1i$delta_0)~1,type='kaplan-meier',
              data=M1i)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5){
  M1i[length(km_c),1:3]<-D3.ind[sample(nrow(D3.ind),1),]
  M1i$X[n]<-min(M1i$T[n],M1i$C[n])
  M1i$delta[n]<-(M1i$X[n] %in% D3.ind$T)*1
  M1i$delta_0[n]<-(M1i$X[n] %in% D3.ind$C)*1
  M1i<-M1i[order(M1i$X),]
  km_c<-survfit(Surv(M1i$X,M1i$delta_0)~1,
                type='kaplan-meier',data=M1i)$surv}

propC31i[k]<-mean((M1i$X %in% D3.ind$C)*1)#Prop. de obs. censuradas
######Estimadores
z<-cbind(rep(1,n),M1i$Y)
km_t<-survfit(Surv(M1i$X,M1i$delta)~1,type='kaplan-meier',
              data=M1i)$surv
Fcn<-sapply(M1i$X,FUN=Fn,M=M1i,k=km_c)
Bs<-Hs(M1i,z,n); km<-exp(-(z%*%Bs)); kzc<-apply(km,2,mean,na.rm=TRUE)
kzc<-kmz(M1i,z,n)
pesos<-M1i$delta/km_c; pesos.z<-M1i$delta/kzc
Fzcn<-sapply(M1i$X,FUN=Fn,M=M1i,k=kzc)
xi<-mean(pesos*(2*Fcn*M1i$X),na.rm=TRUE)
mu<-mean(pesos*M1i$X,na.rm=TRUE)
xiz<-mean(pesos.z*(2*Fzcn*M1i$X),na.rm=TRUE)
muz<-mean(pesos.z*M1i$X,na.rm=TRUE)
#############    RESULTADOS  #################
G31i.Bon[k]<-GBonn(data = cbind(M1i$X,M1i$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1i$X,M1i$delta))
IC.GBon<-c(G31i.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G31i.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G31iIn.Bonn[k]<-(G3i.real>=IC.GBon[1]&G3i.real<=IC.GBon[2])*1

G31i.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1i,km_c,km_t,n,Fcn)
IC.Gesti<-c(G31i.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G31i.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G31iIn.esti[k]<-(G3i.real>=IC.Gesti[1]&G3i.real<=IC.Gesti[2])*1


ginord<-gini(M1i$X)
ginord<-variance(M1i$X,indicator=ginord,R=200,bootType="naive")
G31i.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G31i.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G31i.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G31iIn.ord[k]<-(G3i.real>=IC.Gord[1]&G3i.real<=IC.Gord[2])*1


ginpond<-gini(M1i$X,weights=pesos)
ginpond<-variance(M1i$X,indicator=ginpond,R=200,bootType="naive")
```

```
G31i.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G31i.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G31i.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G31iIn.pond[k]<-(G3i.real>=IC.Gpond[1]&G3i.real<=IC.Gpond[2])*1

G31i.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1i[sample(nrow(M1i),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y)
  Bs<-Hs(Muest,zb,nb); km<-exp(-(zb%*%Bs)); kzc<-apply(km,2,mean,na.rm=TRUE)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G31iIn.dep[k]<-(G3i.real>=IC.Gdep[1]&G3i.real<=IC.Gdep[2])*1

ginpdep<-gini(M1i$X,weights=pesos.z)
ginpdep<-variance(M1i$X,indicator=ginpdep,R=200,bootType="naive")
G31i.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G31i.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G31i.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G31iIn.pdep[k]<-(G3i.real>=IC.Gpdep[1]&G3i.real<=IC.Gpdep[2])*1
}
proc.time()-t

######################################################################
####### SIMULACIÓN PARA CENSURA DEPENDIENTE DE COVARIABLES #######
######################################################################

#### 1. Primer conjunto de datos para censura dependiente #####
#Generación de Población, Valor real y Muestra
Y<-rbinom(1E7,1,0.5)
D1.dep<-data.frame(T=Y*rweibull(1E7,2.3,2)+(1-Y)*rweibull(1E7,1.5,5),
                   C=Y*rweibull(1E7,2.3,3)+(1-Y)*rlnorm(1E7,0,1.9))
D1.dep$Y<-Y
plot(density(D1.dep$T)); plot(density(D1.dep$C)); #lines(density(D1.dep£Y))
G1d.real<-(gini(D1.dep$T)$value)/100#Con paquete "laeken"

##### Ejemplo 1.1.- REPETICIÓN 500 veces con muestra tamaño 300 #####
propC11d<-c(); G11d.Bon<-c(); G11dIn.Bonn<-c(); G11d.esti<-c(); G11dIn.esti<-c()
G11d.ord<-c(); G11dIn.ord<-c(); G11d.pond<-c(); G11dIn.pond<-c(); G11d.dep<-c()
G11dIn.dep<-c(); G11d.pdep<-c(); G11dIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1d<-D1.dep[sample(nrow(D1.dep),3E2),]
M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
M1d$delta<-(M1d$X %in% D1.dep$T)*1
M1d$delta_0<-(M1d$X %in% D1.dep$C)*1
M1d<-M1d[order(M1d$X),]
km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
              data=M1d)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5||length(km_c)<3E2){
  M1d<-D1.dep[sample(nrow(D1.dep),3E2),]
  M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
```

```
  M1d$delta<-(M1d$X %in% D1.dep$T)*1
  M1d$delta_0<-(M1d$X %in% D1.dep$C)*1
  M1d<-M1d[order(M1d$X),]
  km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
                data=M1d)$surv; n<-length(km_c);n ;km_c[n]}

propC11d[k]<-mean((M1d$X %in% D1.dep$C)*1)#Prop. de obs. censuradas
######Estimadores
z<-cbind(rep(1,n),M1d$Y)
km_t<-survfit(Surv(M1d$X,M1d$delta)~1,type='kaplan-meier',
              data=M1d)$surv; length(km_t)
Fcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=km_c)
kzc<-kmz(M1d,z,n)
pesos<-M1d$delta/km_c; pesos.z<-M1d$delta/kzc; mean(pesos.z)
Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xi<-mean(pesos*(2*Fcn*M1d$X),na.rm=TRUE)
mu<-mean(pesos*M1d$X,na.rm=TRUE)
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
#############    RESULTADOS  #################
G11d.Bon[k]<-GBonn(data = cbind(M1d$X,M1d$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1d$X,M1d$delta))
IC.GBon<-c(G11d.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G11d.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G11dIn.Bonn[k]<-(G1d.real>=IC.GBon[1]&G1d.real<=IC.GBon[2])*1

G11d.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1d,km_c,km_t,n,Fcn)
IC.Gesti<-c(G11d.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G11d.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G11dIn.esti[k]<-(G1d.real>=IC.Gesti[1]&G1d.real<=IC.Gesti[2])*1


ginord<-gini(M1d$X)
ginord<-variance(M1d$X,indicator=ginord,R=200,bootType="naive")
G11d.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G11d.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G11d.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G11dIn.ord[k]<-(G1d.real>=IC.Gord[1]&G1d.real<=IC.Gord[2])*1


ginpond<-gini(M1d$X,weights=pesos)
ginpond<-variance(M1d$X,indicator=ginpond,R=200,bootType="naive")
G11d.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G11d.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G11d.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G11dIn.pond[k]<-(G1d.real>=IC.Gpond[1]&G1d.real<=IC.Gpond[2])*1

G11d.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
```

```
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G11dIn.dep[k]<-(G1d.real>=IC.Gdep[1]&G1d.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G11d.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G11d.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G11d.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G11dIn.pdep[k]<-(G1d.real>=IC.Gpdep[1]&G1d.real<=IC.Gpdep[2])*1
}
proc.time()-t

######################################################################
#### 2. Segundo conjunto de datos para censura dependiente ####
#Generación de Población, Valor real y Muestra
Y<-rbinom(1E7,1,0.5)
D2.dep<-data.frame(T=Y*rchisq(1E7,3)+(1-Y)*rchisq(1E7,60),
                   C=Y*rf(1E7,6,3)+(1-Y)*rweibull(1E7,4,70))
D2.dep$Y<-Y
plot(density(D2.dep$T)); plot(density(D2.dep$C)); #lines(density(D1.dep£Y))
G2d.real<-(gini(D2.dep$T)$value)/100#Con paquete "laeken"

##### Ejemplo 2.1.- REPETICIÓN 500 veces con muestra tamaño 300 #####
propC21d<-c(); G21d.Bon<-c(); G21dIn.Bonn<-c(); G21d.esti<-c(); G21dIn.esti<-c()
G21d.ord<-c(); G21dIn.ord<-c(); G21d.pond<-c(); G21dIn.pond<-c(); G21d.dep<-c()
G21dIn.dep<-c(); G21d.pdep<-c(); G21dIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1d<-D2.dep[sample(nrow(D2.dep),3E2),]
M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
M1d$delta<-(M1d$X %in% D2.dep$T)*1
M1d$delta_0<-(M1d$X %in% D2.dep$C)*1
M1d<-M1d[order(M1d$X),]
km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
              data=M1d)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5||length(km_c)<3E2){
  M1d<-D2.dep[sample(nrow(D2.dep),3E2),]
  M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
  M1d$delta<-(M1d$X %in% D2.dep$T)*1
  M1d$delta_0<-(M1d$X %in% D2.dep$C)*1
  M1d<-M1d[order(M1d$X),]
  km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
                data=M1d)$surv; n<-length(km_c);n ;km_c[n]}

propC21d[k]<-mean((M1d$X %in% D2.dep$C)*1)#Prop. de obs. censuradas
######Estimadores
z<-cbind(rep(1,n),M1d$Y)
km_t<-survfit(Surv(M1d$X,M1d$delta)~1,type='kaplan-meier',
              data=M1d)$surv; length(km_t)
Fcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=km_c)
kzc<-kmz(M1d,z,n)
pesos<-M1d$delta/km_c; pesos.z<-M1d$delta/kzc; mean(pesos.z)
Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xi<-mean(pesos*(2*Fcn*M1d$X),na.rm=TRUE)
mu<-mean(pesos*M1d$X,na.rm=TRUE)
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
```

```
#############    RESULTADOS  #################
G21d.Bon[k]<-GBonn(data = cbind(M1d$X,M1d$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1d$X,M1d$delta))
IC.GBon<-c(G21d.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G21d.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G21dIn.Bonn[k]<-(G2d.real>=IC.GBon[1]&G2d.real<=IC.GBon[2])*1

G21d.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1d,km_c,km_t,n,Fcn)
IC.Gesti<-c(G21d.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G21d.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G21dIn.esti[k]<-(G2d.real>=IC.Gesti[1]&G2d.real<=IC.Gesti[2])*1


ginord<-gini(M1d$X)
ginord<-variance(M1d$X,indicator=ginord,R=200,bootType="naive")
G21d.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G21d.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G21d.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G21dIn.ord[k]<-(G2d.real>=IC.Gord[1]&G2d.real<=IC.Gord[2])*1


ginpond<-gini(M1d$X,weights=pesos)
ginpond<-variance(M1d$X,indicator=ginpond,R=200,bootType="naive")
G21d.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G21d.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G21d.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G21dIn.pond[k]<-(G2d.real>=IC.Gpond[1]&G2d.real<=IC.Gpond[2])*1

G21d.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200){
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G21dIn.dep[k]<-(G2d.real>=IC.Gdep[1]&G2d.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G21d.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G21d.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G21d.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G21dIn.pdep[k]<-(G2d.real>=IC.Gpdep[1]&G2d.real<=IC.Gpdep[2])*1
}
proc.time()-t

#########################################################################
#### 3. Tercer conjunto de datos para censura dependiente ####
#Generación de Población, Valor real y Muestra
Y<-rbinom(1E7,1,0.5)
D3.dep<-data.frame(T=Y*rexp(1E7,6)+(1-Y)*rweibull(1E7,2,3),
```

```
                    C=Y*rchisq(1E7,1)+(1-Y)*rweibull(1E7,1,2))
D3.dep$Y<-Y
plot(density(D3.dep$T)); lines(density(D3.dep$C)); #lines(density(D3.dep£Y))
G3d.real<-(gini(D3.dep$T)$value)/100#Con paquete "laeken"

##### Ejemplo 3.1.- REPETICIÓN 500 veces con muestra tamaño 300 #####
propC31d<-c(); G31d.Bon<-c(); G31dIn.Bonn<-c(); G31d.esti<-c(); G31dIn.esti<-c()
G31d.ord<-c(); G31dIn.ord<-c(); G31d.pond<-c(); G31dIn.pond<-c(); G31d.dep<-c()
G31dIn.dep<-c(); G31d.pdep<-c(); G31dIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1d<-D3.dep[sample(nrow(D3.dep),3E2),]
M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
M1d$delta<-(M1d$X %in% D3.dep$T)*1
M1d$delta_0<-(M1d$X %in% D3.dep$C)*1
M1d<-M1d[order(M1d$X),]
km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
              data=M1d)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5||length(km_c)<3E2){
  M1d<-D3.dep[sample(nrow(D3.dep),3E2),]
  M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
  M1d$delta<-(M1d$X %in% D3.dep$T)*1
  M1d$delta_0<-(M1d$X %in% D3.dep$C)*1
  M1d<-M1d[order(M1d$X),]
  km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
                data=M1d)$surv; n<-length(km_c);n ;km_c[n]}

propC31d[k]<-mean((M1d$X %in% D3.dep$C)*1)#Prop. de obs. censuradas
######Estimadores
z<-cbind(rep(1,n),M1d$Y)
km_t<-survfit(Surv(M1d$X,M1d$delta)~1,type='kaplan-meier',
              data=M1d)$surv; length(km_t)
Fcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=km_c)
kzc<-kmz(M1d,z,n)
pesos<-M1d$delta/km_c; pesos.z<-M1d$delta/kzc; mean(pesos.z)
Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xi<-mean(pesos*(2*Fcn*M1d$X),na.rm=TRUE)
mu<-mean(pesos*M1d$X,na.rm=TRUE)
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
#############    RESULTADOS  #################
G31d.Bon[k]<-GBonn(data = cbind(M1d$X,M1d$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1d$X,M1d$delta))
IC.GBon<-c(G31d.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G31d.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G31dIn.Bonn[k]<-(G3d.real>=IC.GBon[1]&G3d.real<=IC.GBon[2])*1

G31d.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1d,km_c,km_t,n,Fcn)
IC.Gesti<-c(G31d.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G31d.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G31dIn.esti[k]<-(G3d.real>=IC.Gesti[1]&G3d.real<=IC.Gesti[2])*1


ginord<-gini(M1d$X)
ginord<-variance(M1d$X,indicator=ginord,R=200,bootType="naive")
G31d.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G31d.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
```

```
           G31d.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G31dIn.ord[k]<-(G3d.real>=IC.Gord[1]&G3d.real<=IC.Gord[2])*1


ginpond<-gini(M1d$X,weights=pesos)
ginpond<-variance(M1d$X,indicator=ginpond,R=200,bootType="naive")
G31d.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G31d.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G31d.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G31dIn.pond[k]<-(G3d.real>=IC.Gpond[1]&G3d.real<=IC.Gpond[2])*1

G31d.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200){
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G31dIn.dep[k]<-(G3d.real>=IC.Gdep[1]&G3d.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G31d.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G31d.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G31d.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G31dIn.pdep[k]<-(G3d.real>=IC.Gpdep[1]&G3d.real<=IC.Gpdep[2])*1
}
proc.time()-t
```

A.3. Análisis de Robustez

```
####################       ANÁLISIS DE ROBUSTEZ       ########################
library(survival)
library(Survgini)
library(laeken)

#Corremos funciones creadas por usuario #

###### Generamos T. #Cáculo de valor real del coeficiente de gini ########
y1<-rchisq(1E7,6);y2<-rweibull(1E7,10,13);u<-runif(1E7);Ti<-y1+(runif(1E7))^2
plot(density(Ti)); G.real<-(gini(Ti)$value)/100; #Con paquete "laeken"


###### 1. PRIMER MECANISMO DE RIESGO  #########
B1<-runif(1E7,0.0,0.04)
E1<-data.frame(T=Ti,C=-(log(u)/(B1*y1)),Y1=y1,Y2=y2); plot(density(E1$C))
plot(density(E1$T));
##### Ejemplo 1.1.- REPETICIÓN 500 veces con muestra tamaño 100 #####
```

```
propC11d<-c(); G11d.Bon<-c(); G11dIn.Bonn<-c(); G11d.esti<-c(); G11dIn.esti<-c()
G11d.ord<-c(); G11dIn.ord<-c(); G11d.pond<-c(); G11dIn.pond<-c(); G11AD.dep<-c()
G11ADIn.dep<-c(); G11AD.pdep<-c(); G11ADIn.pdep<-c(); G11CI.dep<-c()
G11CIIn.dep<-c(); G11CI.pdep<-c(); G11CIIn.pdep<-c(); G11SR.dep<-c()
G11SRIn.dep<-c(); G11SR.pdep<-c(); G11SRIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1d<-E1[sample(nrow(E1),1E2),]
M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
M1d$delta<-(M1d$X %in% E1$T)*1
M1d$delta_0<-(M1d$X %in% E1$C)*1
M1d<-M1d[order(M1d$X),]
km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
              data=M1d)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5||length(km_c)<1E2){
  M1d<-E1[sample(nrow(E1),1E2),]
  M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
  M1d$delta<-(M1d$X %in% E1$T)*1
  M1d$delta_0<-(M1d$X %in% E1$C)*1
  M1d<-M1d[order(M1d$X),]
  km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
                data=M1d)$surv; n<-length(km_c);n ;km_c[n]}

mean((M1d$X %in% E1$C)*1)#Prop. de obs. censuradas propC11d[k]<-
######Estimadores
km_t<-survfit(Surv(M1d$X,M1d$delta)~1,type='kaplan-meier',
              data=M1d)$surv; length(km_t)
Fcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=km_c); pesos<-M1d$delta/km_c
xi<-mean(pesos*(2*Fcn*M1d$X),na.rm=TRUE)
mu<-mean(pesos*M1d$X,na.rm=TRUE)
#############    RESULTADOS  #################
G11d.Bon[k]<-GBonn(data = cbind(M1d$X,M1d$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1d$X,M1d$delta))
IC.GBon<-c(G11d.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G11d.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G11dIn.Bonn[k]<-(G.real>=IC.GBon[1]&G.real<=IC.GBon[2])*1

G11d.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1d,km_c,km_t,n,Fcn)
IC.Gesti<-c(G11d.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G11d.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G11dIn.esti[k]<-(G.real>=IC.Gesti[1]&G.real<=IC.Gesti[2])*1


ginord<-gini(M1d$X)
ginord<-variance(M1d$X,indicator=ginord,R=200,bootType="naive")
G11d.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G11d.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G11d.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G11dIn.ord[k]<-(G.real>=IC.Gord[1]&G.real<=IC.Gord[2])*1


ginpond<-gini(M1d$X,weights=pesos)
ginpond<-variance(M1d$X,indicator=ginpond,R=200,bootType="naive")
G11d.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G11d.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G11d.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
```

```
G11dIn.pond[k]<-(G.real>=IC.Gpond[1]&G.real<=IC.Gpond[2])*1

######################### Modelo Adecuado AD ############################
zAD<-cbind(rep(1,n),M1d$Y1); kzc<-kmz(M1d,zAD,n); pesos.z<-M1d$delta/kzc
mean(pesos.z,na.rm = TRUE); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G11AD.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G11ADIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G11AD.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G11AD.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G11AD.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G11ADIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo CON Variables Irrelevantes CI ########################
zCI<-cbind(rep(1,n),M1d$Y1,M1d$Y2); kzc<-kmz(M1d,zCI,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G11CI.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1,Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G11CIIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G11CI.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G11CI.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G11CI.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G11CIIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo SIN Variables Relevantes SR ##########################
zSR<-cbind(rep(1,n),M1d$Y2); kzc<-kmz(M1d,zSR,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G11SR.dep[k]<-(xiz/muz)-1
```

```
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G11SRIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G11SR.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G11SR.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G11SR.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G11SRIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
}
proc.time()-t


############## 2. SEGUNDO MECANISMO DE RIESGO  ####################
B1<-rweibull(1E7,0.6,0.005)
E2<-data.frame(T=Ti,C=-(log(u)/(B1*(y1^2))),Y1=y1,Y2=y2);plot(density(E2$T))
plot(density(E2$C));
##### Ejemplo 2.1.- REPETICIÓN 500 veces con muestra tamaño 100 #####
propC21d<-c(); G21d.Bon<-c(); G21dIn.Bonn<-c(); G21d.esti<-c(); G21dIn.esti<-c()
G21d.ord<-c(); G21dIn.ord<-c(); G21d.pond<-c(); G21dIn.pond<-c(); G21AD.dep<-c()
G21ADIn.dep<-c(); G21AD.pdep<-c(); G21ADIn.pdep<-c(); G21CI.dep<-c()
G21CIIn.dep<-c(); G21CI.pdep<-c(); G21CIIn.pdep<-c(); G21SR.dep<-c()
G21SRIn.dep<-c(); G21SR.pdep<-c(); G21SRIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1d<-E2[sample(nrow(E2),1E2),]
M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
M1d$delta<-(M1d$X %in% E2$T)*1
M1d$delta_0<-(M1d$X %in% E2$C)*1
M1d<-M1d[order(M1d$X),]
km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
              data=M1d)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5||length(km_c)<1E2){
  M1d<-E2[sample(nrow(E2),1E2),]
  M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
  M1d$delta<-(M1d$X %in% E2$T)*1
  M1d$delta_0<-(M1d$X %in% E2$C)*1
  M1d<-M1d[order(M1d$X),]
  km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
                data=M1d)$surv; n<-length(km_c);n ;km_c[n]}

mean((M1d$X %in% E2$C)*1)#Prop. de obs. censuradas  propC21d[k]<-
######Estimadores
km_t<-survfit(Surv(M1d$X,M1d$delta)~1,type='kaplan-meier',
              data=M1d)$surv; length(km_t)
Fcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=km_c); pesos<-M1d$delta/km_c
xi<-mean(pesos*(2*Fcn*M1d$X),na.rm=TRUE)
mu<-mean(pesos*M1d$X,na.rm=TRUE)
```

```
############    RESULTADOS  #################
G21d.Bon[k]<-GBonn(data = cbind(M1d$X,M1d$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1d$X,M1d$delta))
IC.GBon<-c(G21d.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G21d.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G21dIn.Bonn[k]<-(G.real>=IC.GBon[1]&G.real<=IC.GBon[2])*1

G21d.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1d,km_c,km_t,n,Fcn)
IC.Gesti<-c(G21d.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G21d.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G21dIn.esti[k]<-(G.real>=IC.Gesti[1]&G.real<=IC.Gesti[2])*1


ginord<-gini(M1d$X)
ginord<-variance(M1d$X,indicator=ginord,R=200,bootType="naive")
G21d.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G21d.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G21d.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G21dIn.ord[k]<-(G.real>=IC.Gord[1]&G.real<=IC.Gord[2])*1


ginpond<-gini(M1d$X,weights=pesos)
ginpond<-variance(M1d$X,indicator=ginpond,R=200,bootType="naive")
G21d.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G21d.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G21d.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G21dIn.pond[k]<-(G.real>=IC.Gpond[1]&G.real<=IC.Gpond[2])*1

######################### Modelo Adecuado AD ############################
zAD<-cbind(rep(1,n),M1d$Y1); kzc<-kmz(M1d,zAD,n); pesos.z<-M1d$delta/kzc
mean(pesos.z,na.rm=TRUE); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G21AD.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G21ADIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G21AD.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G21AD.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G21AD.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G21ADIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo CON Variables Irrelevantes CI ########################
zCI<-cbind(rep(1,n),M1d$Y1,M1d$Y2); kzc<-kmz(M1d,zCI,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
```

```
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G21CI.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1,Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G21CIIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G21CI.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G21CI.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G21CI.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G21CIIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo SIN Variables Relevantes SR ########################
zSR<-cbind(rep(1,n),M1d$Y2); kzc<-kmz(M1d,zSR,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G21SR.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G21SRIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G21SR.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G21SR.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G21SR.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G21SRIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
}
proc.time()-t


###### 3. TERCER MECANISMO DE RIESGO  #########
B1<-runif(1E7,0.000,0.002);
E3<-data.frame(T=Ti,C=sqrt(-(4*log(u))/(B1*y1)),Y1=y1,Y2=y2)
plot(density(E3$C),xlim=c(0,100)); plot(density(E3$T))
##### Ejemplo 3.1.- REPETICIÓN 500 veces con muestra tamaño 100 #####
propC31d<-c(); G31d.Bon<-c(); G31dIn.Bonn<-c(); G31d.esti<-c(); G31dIn.esti<-c()
G31d.ord<-c(); G31dIn.ord<-c(); G31d.pond<-c(); G31dIn.pond<-c(); G31AD.dep<-c()
```

```
G31ADIn.dep<-c(); G31AD.pdep<-c(); G31ADIn.pdep<-c(); G31CI.dep<-c()
G31CIIn.dep<-c(); G31CI.pdep<-c(); G31CIIn.pdep<-c(); G31SR.dep<-c()
G31SRIn.dep<-c(); G31SR.pdep<-c(); G31SRIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1d<-E3[sample(nrow(E3),1E2),]
M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
M1d$delta<-(M1d$X %in% E3$T)*1
M1d$delta_0<-(M1d$X %in% E3$C)*1
M1d<-M1d[order(M1d$X),]
km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
              data=M1d)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5||length(km_c)<1E2){
  M1d<-E3[sample(nrow(E3),1E2),]
  M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
  M1d$delta<-(M1d$X %in% E3$T)*1
  M1d$delta_0<-(M1d$X %in% E3$C)*1
  M1d<-M1d[order(M1d$X),]
  km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
                data=M1d)$surv; n<-length(km_c);n ;km_c[n]}

mean((M1d$X %in% E3$C)*1)#Prop. de obs. censuradas propC31d[k]<-
######Estimadores
km_t<-survfit(Surv(M1d$X,M1d$delta)~1,type='kaplan-meier',
              data=M1d)$surv; length(km_t)
Fcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=km_c); pesos<-M1d$delta/km_c
xi<-mean(pesos*(2*Fcn*M1d$X),na.rm=TRUE)
mu<-mean(pesos*M1d$X,na.rm=TRUE)
#############    RESULTADOS  #################
G31d.Bon[k]<-GBonn(data = cbind(M1d$X,M1d$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1d$X,M1d$delta))
IC.GBon<-c(G31d.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G31d.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G31dIn.Bonn[k]<-(G.real>=IC.GBon[1]&G.real<=IC.GBon[2])*1

G31d.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1d,km_c,km_t,n,Fcn)
IC.Gesti<-c(G31d.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G31d.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G31dIn.esti[k]<-(G.real>=IC.Gesti[1]&G.real<=IC.Gesti[2])*1


ginord<-gini(M1d$X)
ginord<-variance(M1d$X,indicator=ginord,R=200,bootType="naive")
G31d.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G31d.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G31d.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G31dIn.ord[k]<-(G.real>=IC.Gord[1]&G.real<=IC.Gord[2])*1


ginpond<-gini(M1d$X,weights=pesos)
ginpond<-variance(M1d$X,indicator=ginpond,R=200,bootType="naive")
G31d.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G31d.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G31d.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G31dIn.pond[k]<-(G.real>=IC.Gpond[1]&G.real<=IC.Gpond[2])*1
```

```
########################## Modelo Adecuado AD ##############################
zAD<-cbind(rep(1,n),M1d$Y1); kzc<-kmz(M1d,zAD,n); pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G31AD.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G31ADIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G31AD.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G31AD.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G31AD.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G31ADIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo CON Variables Irrelevantes CI ########################
zCI<-cbind(rep(1,n),M1d$Y1,M1d$Y2); kzc<-kmz(M1d,zCI,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G31CI.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200){
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1,Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G31CIIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G31CI.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G31CI.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G31CI.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G31CIIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo SIN Variables Relevantes SR ##########################
zSR<-cbind(rep(1,n),M1d$Y2); kzc<-kmz(M1d,zSR,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G31SR.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
```

```
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G31SRIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G31SR.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G31SR.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G31SR.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G31SRIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
}
proc.time()-t


###### 4. CUARTO MECANISMO DE RIESGO  #########
B1<-runif(1E7,0.0,0.0003);
E4<-data.frame(T=Ti,C=sqrt(-(4*log(u))/(B1*(y1^2))),Y1=y1,Y2=y2)
plot(density(E4$C),xlim=c(0,1000)); plot(density(E4$T))
##### Ejemplo 4.1.- REPETICIÓN 500 veces con muestra tamaño 100 #####
propC41d<-c(); G41d.Bon<-c(); G41dIn.Bonn<-c(); G41d.esti<-c(); G41dIn.esti<-c()
G41d.ord<-c(); G41dIn.ord<-c(); G41d.pond<-c(); G41dIn.pond<-c(); G41AD.dep<-c()
G41ADIn.dep<-c(); G41AD.pdep<-c(); G41ADIn.pdep<-c(); G41CI.dep<-c()
G41CIIn.dep<-c(); G41CI.pdep<-c(); G41CIIn.pdep<-c(); G41SR.dep<-c()
G41SRIn.dep<-c(); G41SR.pdep<-c(); G41SRIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1d<-E4[sample(nrow(E4),1E2),]
M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
M1d$delta<-(M1d$X %in% E4$T)*1
M1d$delta_0<-(M1d$X %in% E4$C)*1
M1d<-M1d[order(M1d$X),]
km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
              data=M1d)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5||length(km_c)<1E2){
  M1d<-E4[sample(nrow(E4),1E2),]
  M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
  M1d$delta<-(M1d$X %in% E4$T)*1
  M1d$delta_0<-(M1d$X %in% E4$C)*1
  M1d<-M1d[order(M1d$X),]
  km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
                data=M1d)$surv; n<-length(km_c);n ;km_c[n]}

mean((M1d$X %in% E4$C)*1)#Prop. de obs. censuradas propC41d[k]<-
######Estimadores
km_t<-survfit(Surv(M1d$X,M1d$delta)~1,type='kaplan-meier',
              data=M1d)$surv; length(km_t)
Fcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=km_c); pesos<-M1d$delta/km_c
xi<-mean(pesos*(2*Fcn*M1d$X),na.rm=TRUE)
mu<-mean(pesos*M1d$X,na.rm=TRUE)
#############    RESULTADOS  #################
G41d.Bon[k]<-GBonn(data = cbind(M1d$X,M1d$delta))$GTmax
```

```
V.Bonn<-VGBonn(data = cbind(M1d$X,M1d$delta))
IC.GBon<-c(G41d.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G41d.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G41dIn.Bonn[k]<-(G.real>=IC.GBon[1]&G.real<=IC.GBon[2])*1

G41d.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1d,km_c,km_t,n,Fcn)
IC.Gesti<-c(G41d.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G41d.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G41dIn.esti[k]<-(G.real>=IC.Gesti[1]&G.real<=IC.Gesti[2])*1


ginord<-gini(M1d$X)
ginord<-variance(M1d$X,indicator=ginord,R=200,bootType="naive")
G41d.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G41d.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G41d.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G41dIn.ord[k]<-(G.real>=IC.Gord[1]&G.real<=IC.Gord[2])*1


ginpond<-gini(M1d$X,weights=pesos)
ginpond<-variance(M1d$X,indicator=ginpond,R=200,bootType="naive")
G41d.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G41d.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G41d.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G41dIn.pond[k]<-(G.real>=IC.Gpond[1]&G.real<=IC.Gpond[2])*1

######################### Modelo Adecuado AD ############################
zAD<-cbind(rep(1,n),M1d$Y1); kzc<-kmz(M1d,zAD,n); pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G41AD.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G41ADIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G41AD.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G41AD.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G41AD.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G41ADIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo CON Variables Irrelevantes CI ########################
zCI<-cbind(rep(1,n),M1d$Y1,M1d$Y2); kzc<-kmz(M1d,zCI,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
```

```
G41CI.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200){
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1,Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G41CIIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G41CI.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G41CI.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G41CI.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G41CIIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo SIN Variables Relevantes SR ######################
zSR<-cbind(rep(1,n),M1d$Y2); kzc<-kmz(M1d,zSR,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G41SR.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G41SRIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G41SR.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G41SR.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G41SR.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G41SRIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
}
proc.time()-t



###### 5. QUINTO MECANISMO DE RIESGO  #########
B1<-runif(1E7,0.0,0.1);
E5<-data.frame(T=Ti,C=-log(u)/(0.03*exp(B1*y1)),Y1=y1,Y2=y2)
plot(density(E5$C),xlim=c(0,100)); plot(density(E5$T),xlim=c(0,100))
##### Ejemplo 5.1.- REPETICIÓN 500 veces con muestra tamaño 100 #####
propC51d<-c(); G51d.Bon<-c(); G51dIn.Bonn<-c(); G51d.esti<-c(); G51dIn.esti<-c()
G51d.ord<-c(); G51dIn.ord<-c(); G51d.pond<-c(); G51dIn.pond<-c(); G51AD.dep<-c()
G51ADIn.dep<-c(); G51AD.pdep<-c(); G51ADIn.pdep<-c(); G51CI.dep<-c()
```

```
G51CIIn.dep<-c(); G51CI.pdep<-c(); G51CIIn.pdep<-c(); G51SR.dep<-c()
G51SRIn.dep<-c(); G51SR.pdep<-c(); G51SRIn.pdep<-c()
##### Inicio del Ciclo #####
t <- proc.time()
for (k in 1:500){
M1d<-E5[sample(nrow(E5),1E2),]
M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
M1d$delta<-(M1d$X %in% E5$T)*1
M1d$delta_0<-(M1d$X %in% E5$C)*1
M1d<-M1d[order(M1d$X),]
km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
              data=M1d)$surv; n<-length(km_c);n ;km_c[n]
while(km_c[n]<1E-5||length(km_c)<1E2){
  M1d<-E5[sample(nrow(E5),1E2),]
  M1d$X<-apply(cbind(M1d$T,M1d$C),1,min)
  M1d$delta<-(M1d$X %in% E5$T)*1
  M1d$delta_0<-(M1d$X %in% E5$C)*1
  M1d<-M1d[order(M1d$X),]
  km_c<-survfit(Surv(M1d$X,M1d$delta_0)~1,type='kaplan-meier',
                data=M1d)$surv; n<-length(km_c);n ;km_c[n]}

mean((M1d$X %in% E5$C)*1)#Prop. de obs. censuradas propC51d[k]<-
######Estimadores
km_t<-survfit(Surv(M1d$X,M1d$delta)~1,type='kaplan-meier',
              data=M1d)$surv; length(km_t)
Fcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=km_c); pesos<-M1d$delta/km_c
xi<-mean(pesos*(2*Fcn*M1d$X),na.rm=TRUE)
mu<-mean(pesos*M1d$X,na.rm=TRUE)
##############    RESULTADOS  #################
G51d.Bon[k]<-GBonn(data = cbind(M1d$X,M1d$delta))$GTmax
V.Bonn<-VGBonn(data = cbind(M1d$X,M1d$delta))
IC.GBon<-c(G51d.Bon[k]-qt(0.975,df = (n-1))*sqrt(V.Bonn/n),
           G51d.Bon[k]+qt(0.975,df =(n-1))*sqrt(V.Bonn/n))
G51dIn.Bonn[k]<-(G.real>=IC.GBon[1]&G.real<=IC.GBon[2])*1

G51d.esti[k]<-(xi/mu)-1
V.esti<-VGest(M1d,km_c,km_t,n,Fcn)
IC.Gesti<-c(G51d.esti[k]-qt(0.975,df = (n-1))*sqrt(V.esti/n),
            G51d.esti[k]+qt(0.975,df =(n-1))*sqrt(V.esti/n))
G51dIn.esti[k]<-(G.real>=IC.Gesti[1]&G.real<=IC.Gesti[2])*1


ginord<-gini(M1d$X)
ginord<-variance(M1d$X,indicator=ginord,R=200,bootType="naive")
G51d.ord[k]<-ginord$value/100
V.ord<-ginord$var/100
IC.Gord<-c(G51d.ord[k]-qt(0.975,df=(n-1))*sqrt(V.ord/n),
           G51d.ord[k]+qt(0.975,df=(n-1))*sqrt(V.ord/n))
G51dIn.ord[k]<-(G.real>=IC.Gord[1]&G.real<=IC.Gord[2])*1


ginpond<-gini(M1d$X,weights=pesos)
ginpond<-variance(M1d$X,indicator=ginpond,R=200,bootType="naive")
G51d.pond[k]<-ginpond$value/100
V.pond<-ginpond$var/100
IC.Gpond<-c(G51d.pond[k]-qt(0.975,df=(n-1))*sqrt(V.pond/n),
            G51d.pond[k]+qt(0.975,df=(n-1))*sqrt(V.pond/n))
G51dIn.pond[k]<-(G.real>=IC.Gpond[1]&G.real<=IC.Gpond[2])*1

########################## Modelo Adecuado AD ##############################
```

```
zAD<-cbind(rep(1,n),M1d$Y1); kzc<-kmz(M1d,zAD,n); pesos.z<-M1d$delta/kzc
mean(pesos.z,na.rm=TRUE); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G51AD.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G51ADIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G51AD.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G51AD.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G51AD.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G51ADIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo CON Variables Irrelevantes CI ########################
zCI<-cbind(rep(1,n),M1d$Y1,M1d$Y2); kzc<-kmz(M1d,zCI,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G51CI.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200){
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y1,Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G51CIIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G51CI.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G51CI.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G51CI.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G51CIIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
########## Modelo SIN Variables Relevantes SR ########################
zSR<-cbind(rep(1,n),M1d$Y2); kzc<-kmz(M1d,zSR,n);pesos.z<-M1d$delta/kzc
mean(pesos.z); Fzcn<-sapply(M1d$X,FUN=Fn,M=M1d,k=kzc);
xiz<-mean(pesos.z*(2*Fzcn*M1d$X),na.rm=TRUE)
muz<-mean(pesos.z*M1d$X,na.rm=TRUE)
G51SR.dep[k]<-(xiz/muz)-1
Gdepboot<-c()
for (j in 1:200) {
  Muest<-M1d[sample(nrow(M1d),n*0.9),]
```

```
  Muest<-Muest[order(Muest$X),]
  nb<-length(Muest$X); zb<-cbind(rep(1,nb),Muest$Y2)
  kzc<-kmz(Muest,zb,nb)
  pes.z<-Muest$delta/kzc; Fzcn<-sapply(Muest$X,FUN=Fn,M=Muest,k=kzc)
  xiz<-mean(pes.z*(2*Fzcn*Muest$X),na.rm=TRUE)
  muz<-mean(pes.z*Muest$X,na.rm=TRUE); Gdepboot[j]<-(xiz/muz)-1
}
IC.Gdep<-quantile(Gdepboot,probs = c(0.025,0.975))
G51SRIn.dep[k]<-(G.real>=IC.Gdep[1]&G.real<=IC.Gdep[2])*1

ginpdep<-gini(M1d$X,weights=pesos.z)
ginpdep<-variance(M1d$X,indicator=ginpdep,R=200,bootType="naive")
G51SR.pdep[k]<-ginpdep$value/100
V.pdep<-ginpdep$var/100
IC.Gpdep<-c(G51SR.pdep[k]-qt(0.975,df=(n-1))*sqrt(V.pdep/n),
            G51SR.pdep[k]+qt(0.975,df=(n-1))*sqrt(V.pdep/n))
G51SRIn.pdep[k]<-(G.real>=IC.Gpdep[1]&G.real<=IC.Gpdep[2])*1
}
proc.time()-t
```

Referencias

Alfons, Andreas y Matthias Templ (2013). "Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken". En: *Journal of Statistical Software* 54.15, págs. 1-25. URL: http://www.jstatsoft.org/v54/i15/.

Altman, Douglas G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall/CRC Texts in Statistical Science. Chapman y Hall. URL: https://books.google.com.mx/books?id=5YRpPwAACAAJ.

Andersen, Per Kragh y col. (1993). *Statistical Models Based on Counting Processes*. Springer.

Bain, Lee J. y Max Engelhardt (1992). *Introduction to probability and mathematical statistics*. 2.ª ed. Duxbury/Thomson Learning.

Bonetti, Marco, Chiara Gigliarano y Pietro Muliere (sep. de 2009). "The Gini concentration test for survival data". En: *Lifetime Data Analysis* 15.4, pág. 493. URL: https://doi.org/10.1007/s10985-009-9125-5.

Casella, George y Roger L. Berger (2002). *Statical Inference*. 2.ª ed. Duxbury/Thomson Learning.

Ceriani, Lidia y Paolo Verme (2012). "The origins of the Gini index: extracts from Variabilita e Mutabilita (1912) by Corrado Gini". En: *J Econ Inequal*, págs. 421-443.

Davidson, Russell (mayo de 2009). "Reliable inference for the Gini index". En: *Journal of Econometrics* 150.1, págs. 30-40.

Fleming, Thomas R. y David P. Harrington (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Gastwirth, J. (1972). "The Estimation of the Lorenz Curve and Gini Index". En: *Review of Economics and Statistics*, págs. 306-316.

Gigliarano, Chiara y Marco Bonetti (2011). *Survgini: The Gini concentration test for survival data*. R package version 1.0. URL: https://CRAN.R-project.org/package=Survgini.

Gill, Richard (ene. de 1994). "Glivenko-Cantelli for Kaplan-Meier". En: *Mathematical Methods of Statistics* 3, págs. 76-87.

Guo, S. (2010). *Survival Analysis*. 1.ª ed. Oxford University Press.

Horvitz, D. G. y D. J. Thompson (1952). "A Generalization of Sampling Without Replacement from a Finite Universe". En: *Journal of the American Statistical Association* 47.260, págs. 663-685. URL: http://www.stat.cmu.edu/~brian/905-2008/papers/Horvitz-Thompson-1952-jasa.pdf.

Klein, J. P. y M. L. Moeschberger (2003). *SURVIVAL ANALYSIS: Techniques for Censored and Truncated Data*. 2.ª ed. Springer.

Kleinbaum, D. G. y M. Klein (2012). *Survival Analysis: A Self-Learning Text*. 3.ª ed. Springer Science+Business Media.

Kowalski, Jeanne y Xin M Tu (2008). *Modern applied U-statistics*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley. URL: https://cds.cern.ch/record/1251042.

Lee, Justin (1990). *U-Statistics: Theory and Practice*. Statistics: A Series of Textbooks and Monographs. Taylor y Francis. ISBN: 9780824782535. URL: https://books.google.com.mx/books?id=CZanQod1gSEC.

Lv, Xiaofeng, Gupeng Zhang, Qinghai Li y col. (2016). "Maximum weighted likelihood for discrete choice models with a dependently censored covariate". En: *Journal of the Korean Statistical Society*.

Lv, Xiaofeng, Gupeng Zhang y Guangyu Ren (abr. de 2017). "Gini index estimation for lifetime data". En: *Lifetime Data Analysis: An International Journal Devoted to Statistical Methods and Applications for Time-to-Event Data* 23.2, págs. 275-304.

Martinussen, Torben y Thomas H. Scheike (ene. de 2006). "Dynamic Regression Models for Survival Data". En: *Statistics for Biology and Health*.

O. Aalen, Odd, Ornulf Borgan y Hakon Gjessing (ene. de 2008). *Survival and Event History Analysis: A Process Point of View*. Springer. ISBN: 0-387-20287-0.

Qin, Yongsong, J.N.K. Rao y Changbao Wu (nov. de 2010). "Empirical likelihood confidence intervals for the Gini measure of income inequality". En: *Economic Modelling* 27, págs. 1429-1435.

Robins, James M. y Andrea Rotnitzky (1992). "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers". En: *AIDS Epidemiology: Methodological Issues*. Ed. por Nicholas P. Jewell, Klaus Dietz y Vernon T. Farewell. Boston, MA: Birkhäuser Boston, págs. 297-331. URL: https://doi.org/10.1007/978-1-4757-1229-2_14.

Satten, Glen, Somnath Datta y James Robins (oct. de 2001). "Estimating the marginal survival function in the presence of time dependent covariates". En: *Statistics Probability Letters* 54, págs. 397-403.

Scheike, Thomas H. y Mei-Jie Zhang (2011). "Analyzing Competing Risk Data Using the R timereg Package". En: *Journal of Statistical Software* 38.2, págs. 1-15. URL: http://www.jstatsoft.org/v38/i02/.

Serfling, Robert J. (1980). *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. New York, NY [u.a.]: Wiley.

Zhu, K. (2012). "Comparing Propensity Score And Inverse Weighting Methods In A Longitudinal Time-To-Event Study". En: *Public Health Theses* 1384. URL: https://elischolar.library.yale.edu/cgi/viewcontent.cgi?referer=https://www.google.com.mx/&httpsredir=1&article=1347&context=ysphtdl.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00194
Matrícula: 2173802261

Coeficiente de Gini para datos censurados de duración del desempleo.

En la Ciudad de México, se presentaron a las 13:00 horas del día 18 del mes de diciembre del año 2019 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. JOSE RAUL MONTES DE OCA MACHORRO
DR. HUGO ADAN CRUZ SUAREZ
DRA. BLANCA ROSA PEREZ SALVADOR

Bajo la Presidencia del primero y con carácter de Secretaria la última, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN CIENCIAS (MATEMÁTICAS APLICADAS E INDUSTRIALES)

DE: OSWALDO GUEVARA MUNIVE

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

OSWALDO GUEVARA MUNIVE
ALUMNO

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTE

DR. JOSE RAUL MONTES DE OCA MACHORRO

VOCAL

DR. HUGO ADAN CRUZ SUAREZ

SECRETARIA

DRA. BLANCA ROSA PEREZ SALVADOR