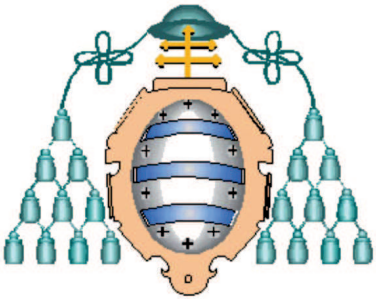
Estimación de la curva ROC acumulativa/dinámica

Sonia Pérez Fernández

Julio de 2015

Estimación de la curva ROC acumulativa/dinámica

Sonia Pérez Fernández



Facultad de Ciencias

Máster Universitario en Modelización e Investigación Matemática, Estadística y Computación

Trabajo Fin de Máster, Oviedo, Julio de 2015

Índice general

Índice de figuras

Introducción

La calidad de una prueba diagnóstica realizada a pacientes no se juzga sólo por sus características analíticas, sino fundamentalmente por su capacidad para distinguir entre diferentes estados de salud. El médico solicita una prueba para decidir, teniendo en cuenta también otros datos disponibles, si el paciente presenta o no una determinada condición clínica. Por lo tanto, para que una prueba se incluya en la práctica clínica habitual, es necesario que ésta sea capaz de reducir la incertidumbre asociada con un cierto estado clínico. La principal cualidad clínica de una prueba diagnóstica es su exactitud, definida como la capacidad para clasificar de manera correcta a los individuos en grupos clínicamente relevantes, como pueden ser dos estados de salud contrarios.

Una forma útil y muy extendida de conocer la calidad de una prueba diagnóstica o de un biomarcador concreto, a lo largo de los posibles puntos de corte elegidos para separar la población sana (casos) de la enferma (controles), es mediante el uso de las denominadas curvas ROC. Pero si el biomarcador que construimos pretende no sólo predecir la aparición o no de una enfermedad o evento (como la muerte), sino también saber cuándo se produce éste, recurrimos a las llamadas curvas ROC tiempo-dependientes, con el fin de analizar su capacidad diagnóstica. En otras palabras, cuando el biomarcador considerado es una variable dependiente del tiempo, la generalización directa es la curva ROC tiempo-dependiente, y particularmente en este trabajo, la curva ROC acumulativa/dinámica. Para un determinado tiempo t, un sujeto se clasifica en el grupo *positivo* (caso) si el evento sucede antes de t, o en el grupo *negativo* (control) si el evento no ocurre hasta después de t.

Ahora bien, en los ensayos clínicos es muy frecuente la aparición de datos censurados,

esto es, por ejemplo, pacientes que han abandonado el estudio por causas desconocidas antes de su fin. La presencia de sujetos censurados, que no pueden ser asignados directamente a uno de los grupos (caso/control) por desconocimiento de cuál ha sido el tiempo real transcurrido entre la inclusión de éstos en el estudio y la ocurrencia del evento, es el principal problema a la hora de estimar este tipo de curvas.

Por ello, el objetivo principal de este trabajo es mostrar una nueva propuesta para la estimación de la curva ROC acumulativa/dinámica, asignando a los sujetos censurados antes del tiempo t una probabilidad de pertenecer al grupo positivo y negativo, respectivamente, en vez de recurrir a alternativas más drásticas como pueden ser la clasificación íntegra en uno de los grupos o la eliminación directa de estos individuos del estudio.

Para facilitar la comprensión del nuevo concepto de curva ROC tiempo-dependiente, y en especial de la curva ROC acumulativa/dinámica, en el primer capítulo se introducen algunas nociones básicas referentes a las curvas ROC y al análisis de supervivencia (ya que incorporamos el parámetro *tiempo*).

Es en el segundo capítulo donde se define la curva ROC tiempo-dependiente y se presentan tres extensiones de ésta según las diferentes definiciones de *casos* y *controles* en el tiempo: incidente/estática, incidente/dinámica y acumulativa/dinámica, siendo esta última la que nos concierne.

El tercer capítulo está dedicado íntegramente a la curva ROC acumulativa/dinámica, mostrando tres posibles estimaciones de ésta junto con las desventajas de cada una de ellas: la primera propuesta está basada en la tradicional función de supervivencia de Kaplan-Meier, la segunda en el estimador de vecinos próximos (KNN) para una distribución bivariada bajo censura aleatoria y el tercer estimador se basa en la curva de incidencia acumulativa (CIC) propuesta por Nelson-Aalen. La parte principal de este capítulo, y en general de este trabajo, es la inclusión de una cuarta propuesta, la cual pretende servir como solución a los problemas que presentan las anteriores.

Finalmente, en los capítulos cuarto y quinto, se muestra el comportamiento de todos los posibles estimadores mencionados en el capítulo anterior, tanto con el uso de datos simulados, como utilizando una base de datos real.

Capítulo 1

Preliminares

En este capítulo se presentarán algunas nociones básicas necesarias para la posterior comprensión del concepto de curva ROC tiempo-dependiente. Así, en primer lugar se introducirá el concepto de curva ROC, junto con otras definiciones que la acompañan por naturaleza, como son la noción de *especificidad* y de *sensibilidad*. En segundo lugar, se expondrán asimismo algunas nociones propias del Análisis de Supervivencia, haciendo especial hincapié en el tratamiento de datos censurados. Además, para facilitar la comprensión del método de estimación propuesto como objetivo final de este trabajo, se introducen algunas pinceladas acerca del estimador de Kaplan-Meier y de la regresión de Cox.

1.1. Nociones básicas de la curva ROC

1.1.1. Un poco de historia

La curva ROC (del inglés *Receiver Operating Characteristics*) es una técnica para visualizar, organizar y seleccionar clasificadores según su comportamiento. Las curvas ROC fueron utilizadas en sus inicios, a mediados de los años cincuenta, en teoría de detección de señales [9], y posteriormente fueron extendidas para visualizar y analizar el comportamiento de sistemas diagnósticos. A día de hoy, existe una amplia literatura

acerca del uso de este tipo de gráficos para los test diagnósticos. En el año 2000, el artículo de *Swets et al.* [32] captó la atención de muchas personas, animando a la utilización de esta potente herramienta.

Uno de los primeros en utilizar las curvas ROC en Aprendizaje Automático (en inglés *Machine Learning*) fue *Spackman* en 1989 [31], el cual demostró el poder de estas curvas para evaluar y comparar algoritmos. Es notorio el gran incremento en el uso de curvas ROC en este campo durante los últimos años, debido en parte a que el cálculo de una simple medida de precisión de un clasificador es, normalmente, una evaluación pobre para medir su comportamiento. Además de ser un método gráfico de uso muy generalizado y útil, es conceptualmente simple, lo cual facilita su manejo.

En medicina, una de las primeras aplicaciones de la curva ROC fue publicada en los años sesenta por *Lusted* [17], aunque ésta ganó más popularidad en la década de los setenta (*Martinez et al., 2003* [19]; *Zhou et al., 2011* [39]). A día de hoy, las nuevas tecnologías en el ámbito clínico ofrecen un amplio abanico de posibilidades para diagnosticar una enfermedad, o predecir la progresión de ésta, y por ello los test/pruebas diagnósticas y los biomarcadores están en continuo estudio. El análisis de la curva ROC se suele utilizar para evaluar el poder discriminatorio (de clasificar en clases contrarias) de una variable continua que representa un test diagnóstico, un biomarcador o un clasificador.

Este tipo de gráficos son útiles para la consecución de diversos objetivos:

- Evaluar la capacidad de discriminación de un marcador continuo para asignar correctamente los individuos a dos grupos distintos.

- Buscar el punto de corte óptimo para minimizar la mala clasificación de estos sujetos.

- Comparar la eficacia de dos o más marcadores o pruebas diagnósticas.

- Estudiar la variabilidad entre observadores cuando dos o más miembros de la comunidad científica miden la misma variable continua.

1.1.2. Sensibilidad y especificidad

Tomamos X una variable aleatoria continua (aunque en el caso discreto sería similar) que recoge la medición resultante de una prueba diagnóstica, o más comúnmente, biomarcador, medida tanto en individuos sanos ($D = 0$ si denotamos por D el verdadero estado de salud del sujeto) como enfermos ($D = 1$). Vamos a suponer, sin pérdida de generalidad, que para un determinado **punto de corte** x_0, el resultado del test es positivo (se considera que el individuo está enfermo) si X es mayor que x_0, y negativo en caso contrario.

Al comparar los resultados de la prueba a evaluar (X) y el diagnóstico de referencia (D), existen cuatro posibilidades que pueden resumirse en una tabla de contingencia 2×2 como la que se muestra a continuación:

		Diagnóstico verdadero		
		Caso ($D = 1$)	**Control** ($D = 0$)	
Test diagnóstico	**Positivo** ($X > x_0$)	TP (nº de verdaderos positivos)	FP (nº de falsos positivos)	$TP + FP$ (nº de resultados positivos)
	Negativo ($X \leq x_0$)	FN (nº de falsos negativos)	TN (nº de verdaderos negativos)	$FN + TN$ (nº de resultados negativos)
		$TP + FN$ (nº de casos)	$FP + TN$ (nº de controles)	

la cual da lugar a la siguiente tabla:

		Diagnóstico verdadero	
		Caso ($D = 1$)	**Control** ($D = 0$)
Test diagnóstico	**Positivo** ($X > x_0$)	$TPR = P(X > x_0 \mid D = 1) \simeq \frac{TP}{TP + FN}$ (proporción de verdaderos positivos)	$FPR = P(X > x_0 \mid D = 0) \simeq \frac{FP}{FP + TN}$ (proporción de falsos positivos)
	Negativo ($X \leq x_0$)	$FNR = P(X \leq x_0 \mid D = 1) \simeq \frac{FN}{TP + FN}$ (proporción de falsos negativos)	$TNR = P(X \leq x_0 \mid D = 0) \simeq \frac{TN}{FP + TN}$ (proporción de verdaderos negativos)

Utilizando esta notación, se definen de forma natural los conceptos de **sensibilidad** y **especificidad** como sigue:

Definición 1.1.1. *La **sensibilidad** de un biomarcador continuo X se define como la probabilidad de obtener un resultado positivo cuando el individuo tiene la enfermedad. Es decir, mide su capacidad para detectar la enfermedad cuando ésta está presente. Se denota por S_E y es igual a*

$$S_E(x_0) = TPR = P(X > x_0 \mid D = 1) \, .$$

Si X es nominal y toma más de dos categorías, se puede definir la sensibilidad de manera equivalente como

$$S_E = P(X \in C_1 \mid D = 1)$$

donde C_1 son las clases asociadas a un diagnóstico positivo (caso).

Definición 1.1.2. *La **especificidad** de un biomarcador continuo X se define como la probabilidad de obtener un resultado negativo cuando el individuo no tiene la enfermedad. Es decir, mide su capacidad para descartar la enfermedad cuando ésta no está presente. Se denota por S_P y es igual a*

$$S_P(x_0) = TNR = P(X \leq x_0 \mid D = 0) \, .$$

Si X es nominal y toma más de dos categorías, se puede definir la especificidad de manera equivalente como

$$S_P = P(X \in C_0 \mid D = 0)$$

donde C_0 son las clases asociadas a un diagnóstico negativo (control).

Cabe notar que la sensibilidad se obtiene en el subgrupo de enfermos ($D = 1$) y la especificidad en el de sanos ($D = 0$), por lo que ambos valores son independientes de la prevalencia de la enfermedad en la muestra estudiada.

La prueba diagnóstica ideal sería aquella con una sensibilidad y especificidad próximas a 1, esto es, aquella cuya probabilidad para clasificar correctamente a los individuos, en sanos o enfermos según corresponda, sea muy alta.

1.1.3. La curva ROC

Una curva ROC es un gráfico bidimensional que representa la proporción de verdaderos positivos frente a la proporción de falsos positivos asociados a un biomarcador a lo largo de los distintos puntos de corte que puede tomar éste.

Por tanto, en el eje de ordenadas se muestran los distintos valores de la *sensibilidad*, o lo que es lo mismo TPR, mientras que en el eje de abscisas se presentan los distintos valores de *1 - especificidad* , o lo que es lo mismo FPR.

Formalmente, la curva ROC se definiría como sigue:

Dada una determinada medida proporcionada por un biomarcador X realizada sobre una población de positivos X_P, y otra de negativos X_N, con función de distribución F_P y F_N respectivamente. Suponiendo que $E(X_N) \leq E(X_P)$, para clasificar a los individuos en uno u otro grupo se debe fijar un criterio, punto de corte, a partir del cual será considerado positivo. Por tanto, fijado un punto de corte x_0, la sensibilidad de la prueba vendrá determinada por $1 - F_P(x_0)$ $(= 1 - P(X_P > x_0))$, siendo $F_N(x_0)$ $(= P(X_N \leq x_0))$ su especificidad y quedando por tanto determinada la curva ROC por las coordenadas del vector $(1 - F_N(x_0),\, 1 - F_P(x_0))$ para $x_0 \in \mathbb{R}$, o, equivalentemente, por la función que a cada $p \in [0, 1]$ le asocia

$$\mathcal{R}(p) = 1 - F_P\left[(1 - F_N)^{-1}(p)\right] = 1 - P(X_P \leq (1 - F_N)^{-1}(p)) = P(X_P > (1 - F_N)^{-1}(p))$$

y por las propiedades de la función de distribución empírica,

$$\mathcal{R}(p) = P(X_P > F_N^{-1}(1-p)) = P(F_N(X_P) \geq 1-p) = P(1-F_N(X_P) \leq p) = F_{1-F_N(X_P)}(p)$$

donde $F_{1-F_N(X_P)}$ denota la función de distribución de la variable aleatoria $1 - F_N(X_P)$.

Como siempre, el problema surge cuando no se conocen las distribuciones reales de la variable X en las poblaciones de positivos y negativos y, a partir de sendas muestras aleatorias, deben estimarse. Una de las posibilidades es suponer que las poblaciones siguen algun modelo paramétrico, el gaussiano usualmente, o bien, aplicar algun método no

paramétrico, siendo los más frecuentes sustituir las funciones de distribución desconocidas por sus Funciones de Distribución Empíricas (FDE) o por las Funciones de Distribución Empíricas Suavizadas (FDES).

Si no se hace ninguna suposición sobre la distribución de las variables, el método más frecuente de estimación consiste en sustituir las funciones de distribución desconocidas por sus correspondientes funciones de distribución empíricas. Se tiene que, dadas muestras de positivos X_P y de negativos X_N de tamaños m y n y distribuciones F y G respectivamente, la estimación empírica para la curva ROC viene dada por,

$$\widehat{R}(p) = 1 - \hat{F}_m \left[(1 - \hat{G}_n)^{-1}(p) \right]$$

donde $\hat{F}_m$ es la funcion de distribución empírica asociada a la muestra X_P y $(1 - \hat{G}_n)^{-1}(p) = inf\{x : (1 - \hat{G}_n)(x) \leq p\}$ siendo $\hat{G}_n$ la función de distribución empírica asociada a la muestra X_N.

Para ilustrar esta definición se muestra un ejemplo en el que se asume que X_P y X_N siguen ambas distribuciones normales, con $E(X_P) > E(X_N)$. Se ha considerado que $X_N \equiv \mathcal{N}(-1, 3)$ y $X_P \equiv \mathcal{N}(3, 4)$.

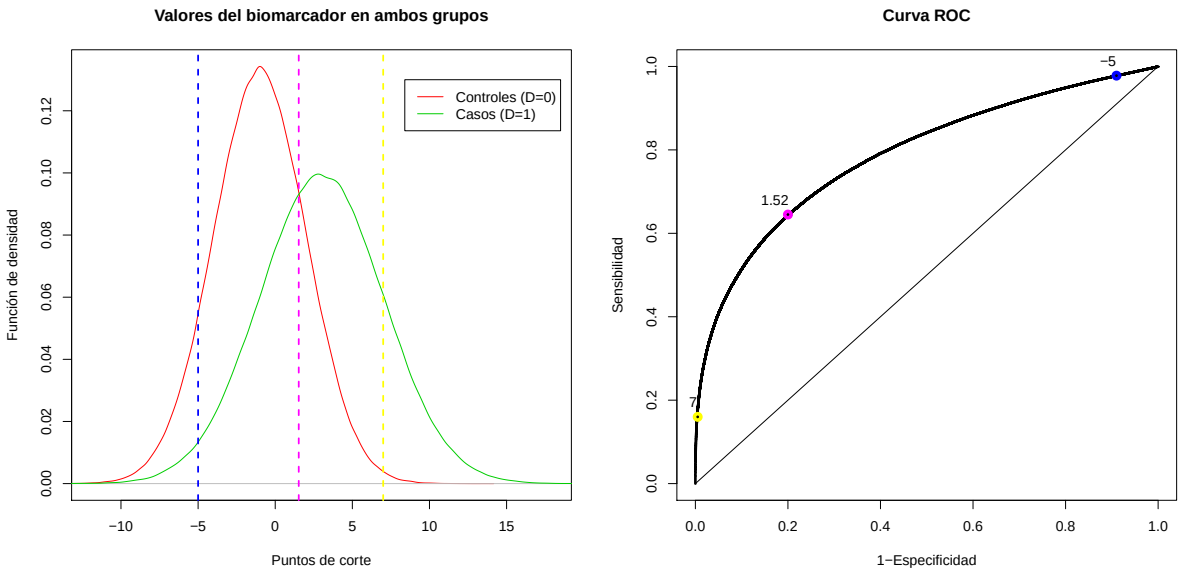


Figura 1.1: Ejemplo Curva ROC

En la figura anterior se muestran los valores de un biomarcador medido tanto en individuos sanos (curva roja) como enfermos (curva verde). En el gráfico de la izquierda se observa que ambas distribuciones se solapan, por lo que sea cual sea el punto de corte que tomemos, no todos los individuos estarán bien clasificados, como es habitual en la práctica. Se han tomado tres puntos de corte diferentes (-5, -1.52 por ser el punto en el que se cortan ambas funciones de densidad y 7). En la gráfica de la derecha se observa la curva ROC resultante de este marcador, en la que aparecen señalados los valores que toma ésta para los puntos de corte anteriores.

Se observa que la *sensibilidad*, tomando como punto de corte 7, es muy pequeña (cerca del 20 %), mientras que tomando el punto -5 es muy grande (cerca del 95 %). Esto es debido a que al tomar el -5 (línea azul en la figura de la izquierda), el área que queda a la izquierda de éste por debajo de la curva verde es muy pequeña, esto es, la proporción de casos mal clasificados ($FNR = 1 - sensibilidad$).

Asimismo, se observa que la *especificidad*, tomando como punto de corte 7, es muy grande (cerca del 100 %), mientras que tomando el punto -5 es muy pequeña (cerca del 0.05 %). Esto es debido a que al tomar el 7 (línea amarilla en la figura de la izquierda), el área que queda a la derecha de éste por debajo de la curva roja es muy pequeña, esto es, la proporción de controles mal clasificados ($FPR = 1 - especificidad$).

El punto de corte que separa ambas funciones de densidad, por su parte, parece que equilibra en gran medida la sensibilidad y la especificidad, siendo la primera de 0.63 y la segunda de 0.78, aproximadamente. Efectivamente, en el gráfico de la izquierda se puede comprobar que la proporción de casos mal clasificados (área encerrada a la izquierda de la línea violeta y por debajo de la función de densidad verde) y la proporción de controles mal clasificados (área encerrada a la derecha de la línea violeta y por debajo de la función de densidad roja) son similares.

Se concluye por tanto que:

- Tomando como punto de corte -5, la prueba es muy sensible, por lo que daría más importancia al hecho de clasificar correctamente a los individuos enfermos, permitiendo sin embargo que muchos sujetos que no tengan la enfermedad sean

clasificados igualmente como enfermos (baja especificidad).

- Tomando como punto de corte 7, la prueba es muy específica, por lo que daría más importancia al hecho de clasificar correctamente a los individuos sanos, permitiendo sin embargo que muchos sujetos que realmente tienen la enfermedad sean clasificados igualmente como sanos (baja sensibilidad).

- Tomando como punto de corte 1.52, la prueba es equilibrada, dando casi la misma importancia al hecho de clasificar correctamente a los individuos enfermos como a los sanos, aun a sabiendas de que estas probabilidades de buena clasificación no son tan altas como las alcanzadas en los casos anteriores.

1.1.4. Medidas del comportamiento del biomarcador

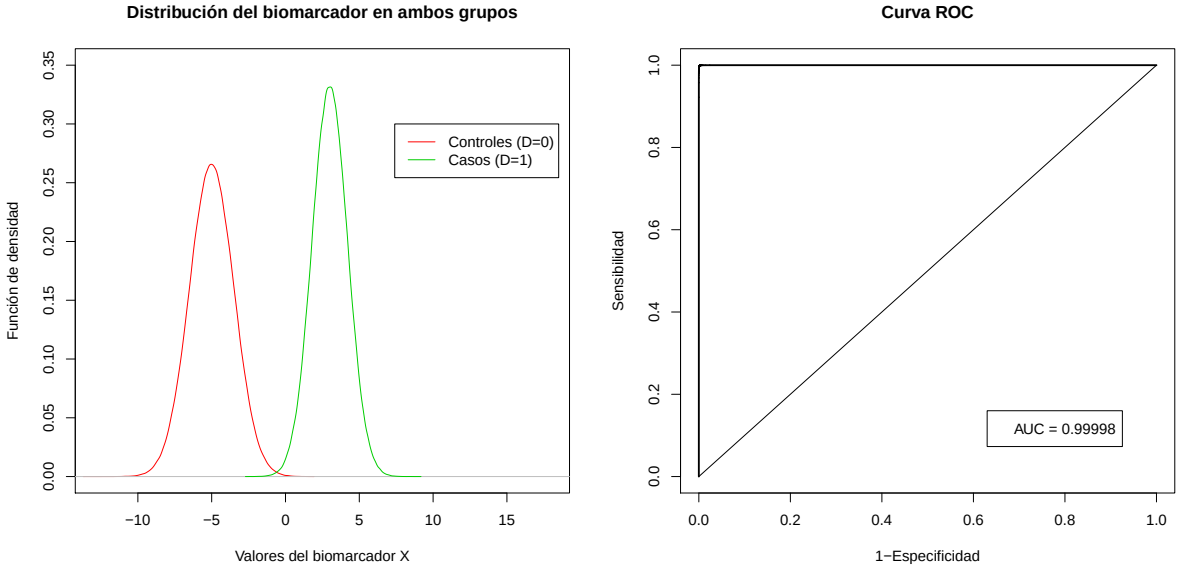
Tras estudiar algunos de los resultados gráficamente visibles, una pregunta evidente queda en el aire: *¿Cómo saber cuán bueno es este biomarcador para clasificar correctamente a la población?*

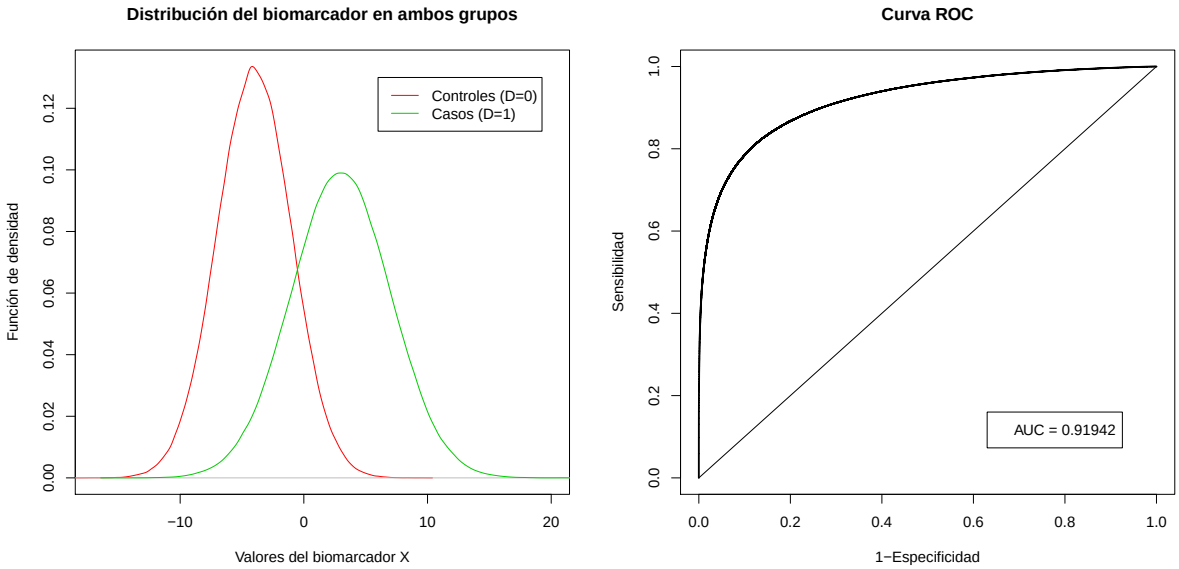
Las curvas ROC no sólo aportan información cualitativa, sino que también permiten la realización de análisis estadísticos para evaluar cuantitativamente las pruebas diagnósticas, teniendo en cuenta que la variabilidad del muestreo puede dar lugar a distintos valores de sensibilidad y especificidad para un mismo punto de corte. Esta variabilidad se mide calculando los intervalos de confianza de la curva ROC en todos los puntos.

Una herramienta muy utilizada para medir la exactitud de un biomarcador es el área bajo la curva ROC, *AUC* (del inglés *Area Under Curve*), que se define como la probabilidad de clasificar correctamente un par de individuos sano y enfermo, seleccionados al azar de la población, mediante el biomarcador considerado. El AUC toma un valor entre 0 y 1, siendo mejor el comportamiento del biomarcador cuanto más cercano a 1 sea este valor. En la práctica, el límite inferior para el área bajo la curva ROC es 0.5, el cual corresponde al área encerrada bajo el segmento diagonal representado en la *Figura 1.1*, que sería la curva ROC resultante de un biomarcador basado en el puro azar, sin ninguna habilidad

intrínseca para discriminar entre sujetos con y sin determinada enfermedad. Cabe notar que como mencionábamos anteriormente, los valores de *sensibilidad* y *especificidad* no dependen de la prevalencia de la enfermedad (diferencias de tamaño muestral entre ambos grupos), por lo que el AUC también es independiente de ésta.

Veamos a continuación varias curvas ROC con sus correspondientes áreas, para observar cómo se pueden comparar varios biomarcadores entre sí, y cuál es el motivo de la forma que toman estas curvas.





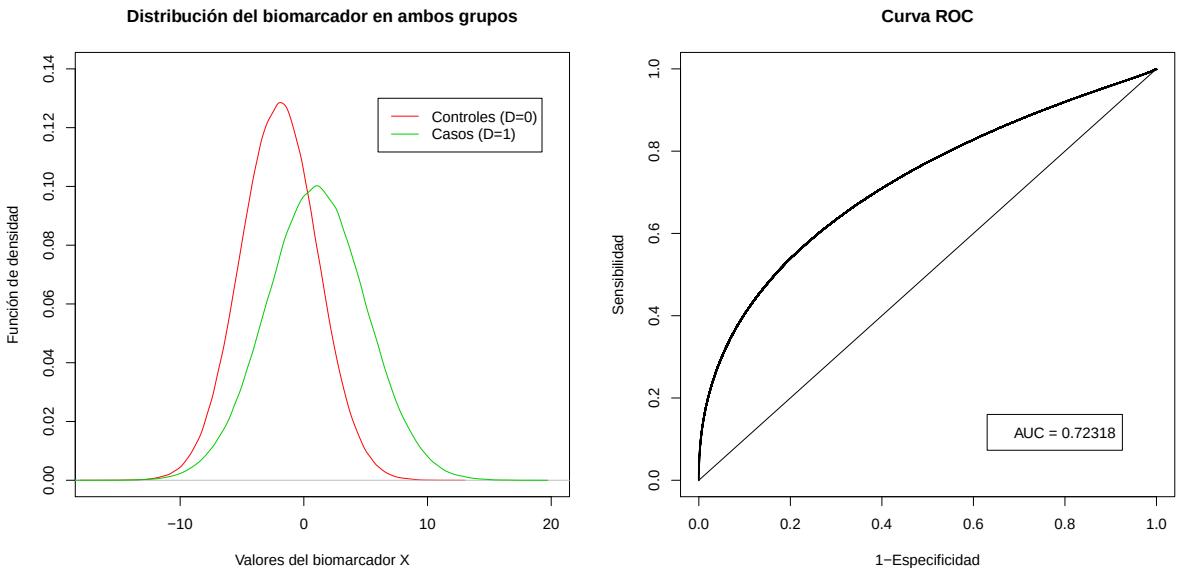


Figura 1.2: Curvas ROC para distintos biomarcadores cuya distribución en ambos subgrupos se solapa cada vez más

Como se puede observar en la *Figura 1.2*, cuanto menos se solapan los valores del biomarcador en los subgrupos de casos y controles (primera gráfica), mayor es el área bajo la curva ROC y, por tanto, mayor es la capacidad clasificatoria del biomarcador. Sin embargo, cuanto más se solapan estos valores (última gráfica), más próxima está la curva ROC a la diagonal, por lo que menor es el AUC, lo que significa que la capacidad clasificatoria del biomarcador es más baja.

Cabe mencionar que una curva ROC por debajo de la diagonal sería consecuencia de la clasificación inversa del biomarcador, por lo que bastaría con reformular éste para tener la representación gráfica correcta. Esto es, si el biomarcador correspondiente toma valores más pequeños en los individuos enfermos que en los sanos (al contrario de como estamos considerando), se debe clasificar el resultado del test diagnóstico, teniendo en cuenta este hecho, de la siguiente manera: si $X > x_0$ el test es negativo (clasifica el individuo correspondiente como sano), mientras que si $X \leq x_0$ lo clasifica como enfermo.

1.2. Nociones básicas del análisis de supervivencia

1.2.1. Introducción al análisis de supervivencia

El objetivo del Análisis de Supervivencia es estudiar el tiempo que transcurre desde la ocurrencia de un determinado suceso (el comienzo de un tratamiento, diagnóstico de un cáncer, un transplante, etc.) hasta la ocurrencia de otro (curación de la enfermedad, muerte, etc.).

El seguimiento viene definido por una fecha de inicio y una fecha de fin (ocurrencia del evento o censura, de la cual hablaremos posteriormente), que determinan el **tiempo de seguimiento**, el cual será denotado de aquí en adelante por la variable aleatoria T. Ambas fechas son diferentes para cada individuo, pues los pacientes se incorporan y "finalizan" en momentos diferentes.

En las enfermedades crónicas, tales como el cáncer, la **supervivencia** se mide como una probabilidad de permanecer vivo durante una determinada cantidad de tiempo. Por ejemplo, el pronóstico del cáncer se valora en función del porcentaje de pacientes que sobreviven al menos cinco años después del diagnóstico.

Con las técnicas del análisis de supervivencia podemos:

- Conocer la probabilidad de sobrevivir a lo largo del tiempo ante la presencia de una enfermedad, transplante, etc.
- Estimar las tasas de supervivencia en una población en función de ciertos factores.
- Comparar estadísticamente la eficacia de distintos tratamientos sobre dicha supervivencia.

1.2.2. La función de supervivencia y la función de riesgo

Definición 1.2.1. *Sea T el tiempo en que ocurre el suceso de interés en el estudio, se define la* ***función de supervivencia*** *como la probabilidad de que el suceso suceda después del tiempo t, o lo que es lo mismo, de sobrevivir al menos un tiempo t. Se denota por $S(t)$*

y se calcula como

$$S(t) = P(T > t) = 1 - F_T(t) \ .$$

Por cómo está definida la función de supervivencia y teniendo en cuenta que, por naturaleza, el tiempo de supervivencia T es no negativo, sabemos que $S(t)$ es una función decreciente tal que $S(0) = 1$ y $\lim_{t\to\infty} S(t) = 0$, esto es, la probabilidad de sobrevivir al menos al tiempo cero es 1, mientras que la de sobrevivir un tiempo infinito es 0. Además, como consecuencia inmediata de la definición, se tiene que

$$S(t) = P(T > t) = \int_t^\infty f_T(u)du \ ,$$

donde f_T denota la función de densidad de la variable T, y además,

$$S'(t) = -F'_T(t) = -f_T(t) \ .$$

La tasa de decrecimiento de una curva de supervivencia varía según el riesgo que tenga el evento de suceder en el tiempo t. De aquí surge la definición de **función de riesgo**:

Definición 1.2.2. *La **función de riesgo** se define como la probabilidad de ocurrencia del evento durante un intervalo de tiempo muy pequeño, suponiendo que el sujeto en estudio ha sobrevivido hasta el tiempo t. Es decir, es el límite de la probabilidad de que un sujeto presente el suceso en el siguiente instante de tiempo, esto es, en un intervalo muy corto, de t a $t + \Delta t$, dado que el individuo ha sobrevivido hasta el inicio del intervalo (tiempo t). Se denota por $h(t)$ y queda definida por la siguiente expresión*

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

En caso de que T sea una variable aleatoria continua, se cumple que

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

Demostración.

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

$$= \lim_{\Delta t \to 0^+} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \cdot \frac{1}{P(T > t)} = \frac{1}{P(T > t)} \cdot \lim_{\Delta t \to 0^+} \frac{P(t < T \leq t + \Delta t)}{\Delta t}$$

$$= \frac{1}{S(t)} \cdot f(t) = \frac{f(t)}{S(t)} \, .$$

□

Cabe destacar que si la función de riesgo es constante, como sucede en el modelo exponencial (en el que la función de supervivencia $S(t)$ es igual a $e^{-\lambda t}$), la probabilidad de que se presente el evento en el siguiente instante es independiente del tiempo. Sin embargo, en muchos problemas reales esta probabilidad varía con el tiempo.

Estrechamente ligado a esta definición surge el concepto de **riesgo acumulado**, que se define como sigue:

Definición 1.2.3. *La **función de riesgo acumulado** se define como la acumulación del riesgo al paso del tiempo. Se denota por $H(t)$ y, en el caso de que T sea una variable aleatoria continua, viene dada por la siguiente expresión*

$$H(t) = \int_0^t h(u)du = -log(S(t)) \, .$$

Una consecuencia inmediata de la definición es

$$S(t) = e^{-H(t)} \, ,$$

por lo que si se conoce la función de riesgo o función de riesgo acumulado, se conoce la función de supervivencia, y viceversa.

Cabe notar además que se trata de una función no decreciente, y según cuál sea su incremento, se puede tener información acerca del comportamiento del riesgo a lo largo del tiempo, lo cual es una ventaja en el análisis de supervivencia.

Las **curvas de supervivencia** representan la tasa o proporción de supervivencia en función del tiempo. Pueden ser estimadas tanto por métodos paramétricos, si se conoce la familia de funciones a la que pertenece la función de supervivencia, o mediante métodos no paramétricos, a través de funciones escalonadas con saltos en los tiempos de muerte de los pacientes observados.

Modelos paramétricos comunes

Algunas funciones de supervivencia pueden ser caracterizadas por familias de distribuciones específicas que sólo dependen de uno o varios parámetros desconocidos, los cuales proporcionan las características específicas del modelo en estudio. La selección de un modelo paramétrico se hace usualmente mediante la función de riesgo, ya que de acuerdo a la información que el investigador tenga del fenómeno que causa el evento, puede determinar las características que el modelo debe seguir en cuanto a la forma de la tasa de riesgo conforme avanza el tiempo. Por ejemplo, puede que el riesgo de muerte de un paciente después de someterse a una cirugía sea creciente las primeras horas y después, si sobrevive, su salud se estabilice hasta lograr su recuperación. En este caso, una función de riesgo creciente en valores pequeños de tiempo, que alcance un máximo y luego sea decreciente, puede ser conveniente para modelar este fenómeno.

Utilizar un modelo paramétrico es restrictivo en el sentido de que se pueden exigir formas específicas del riesgo en el tiempo. Por ejemplo, el modelo exponencial, que presenta riesgo constante, resultaría inadecuado para modelar el tiempo que tarda un individuo en morir cuando se le ha detectado una enfermedad terminal, ya que en este caso, el riesgo debe ser claramente creciente.

A continuación se presentan las distribuciones más comunes en modelos de supervivencia y una explicación detallada de la forma de su función de riesgo, ya que ésta es muy importante a la hora de seleccionar el modelo. Además de las tratados en las líneas sucesivas, existen muchos otros modelos, como el *modelo Erlang*, el *modelo Log-logístico* y el *modelo Pareto*.

Modelo Exponencial

Su función de supervivencia está dada por $S(t) = e^{-\lambda t}$ con $\lambda > 0$.

Su función de densidad es $f(t) = \lambda e^{-\lambda t}$ y está caracterizada por su función de riesgo constante, $h(t) = \lambda$.

La distribución exponencial tiene la propiedad de pérdida de memoria, esto es

$$P(T \geq t + z \mid T \geq t) = P(T \geq z) \, ,$$

de la cual se sigue que la vida media residual en el tiempo t, $E(T - t \mid T > t)$ es igual a $E(T) = 1/\lambda$.

Así, el tiempo de ocurrencia de un evento no depende de lo que haya sucedido en el pasado. La propiedad de pérdida de memoria también es reflejada en la interpretación de riesgo constante, donde la probabilidad de ocurrencia del evento en un instante t, dado que éste no ha ocurrido antes, es independiente de t.

Puesto que la distribución exponencial es un caso particular de las distribuciones *Weibull* y *Gamma*, hereda propiedades de éstas.

Modelo Weibull

Su función de supervivencia está dada por $S(t) = e^{-\lambda t^{\alpha}}$ con $\lambda > 0$ (parámetro de escala) y $\alpha > 0$ (parámetro de forma).

Su función de densidad es $f(t) = \alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}}$ y su función de riesgo está dada por $h(t) = \alpha\lambda t^{\alpha-1}$.

Esta función de riesgo es creciente si $\alpha > 1$, decreciente si $\alpha < 1$ y constante si $\alpha = 1$ (modelo exponencial), lo cual permite modelar el tiempo de ocurrencia del evento para distintas tasas de riesgo a través del tiempo.

Modelo Log-normal

Se dice que la distribución de una variable aleatoria T es log-normal, cuando su logaritmo neperiano, $Y = log(T)$, sigue una distribución normal.

Su función de densidad queda completamente especificada por los parámetros μ y σ, los cuales corresponden a la media y varianza de Y, y viene dada por la siguiente expresión

$$f(t) = \frac{exp\left(-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right)}{t\sqrt{2\pi}\sigma} = \frac{1}{t}\cdot\phi\left(\frac{log(t)-\mu}{\sigma}\right)$$

donde ϕ denota la función de densidad de una normal estándar, $\mu \in \mathbb{R}$ y $\sigma \in (0, \infty)$.

La función de supervivencia está dada por

$$S(t) = 1 - \Phi\left(\frac{log(t)-\mu}{\sigma}\right) ,$$

donde Φ denota la función de distribución de una normal estándar.

La función de riesgo de la distribución log-normal $\big(h(t) = f(t)/S(t)$ con $f(t)$ y $S(t)$ anteriores$\big)$ tiene una forma de "joroba", ya que toma el valor cero en el tiempo cero, después crece a un máximo y decrece a cero en el límite (cuando $t \to \infty$). La crítica usual hacia esta distribución viene infundada por el hecho de ser decreciente para valores grandes de t, lo cual no tiene sentido en muchas ocasiones. Sin embargo, este modelo puede ser útil cuando no interesan valores grandes del tiempo.

Modelo Gamma

La distribución *Gamma* tiene propiedades muy parecidas a las de la distribución *Weibull*, pero ésta es más difícil de tratar matemáticamente. Su función de densidad viene dada por

$$f(t) = \frac{\lambda^\beta t^{\beta-1} e^{-\lambda t}}{\Gamma(\beta)}$$

donde $\lambda > 0$ (parámetro de escala), $\beta > 0$ (parámetro de forma) y $\Gamma(\beta)$ es la función gamma, esto es, $\Gamma(\beta) = \displaystyle\int_0^\infty x^{\beta-1}e^{-x}\,dx$.

Esta distribución se corresponde con el modelo exponencial cuando $\beta = 1$, mientras que cuando $\beta \to \infty$ se aproxima a una distribución normal.

La función de supervivencia está dada por

$$S(t) = \frac{\int_t^\infty \lambda(\lambda x)^{\beta-1}e^{-\lambda x}dx}{\Gamma(\beta)} = 1 - \frac{\int_0^{\lambda t} x^{\beta-1}e^{-x}dx}{\Gamma(\beta)} = 1 - \Gamma^*(\lambda t, \beta)$$

donde Γ^* es la función gamma incompleta $\Gamma^*(t,\beta) = \dfrac{1}{\Gamma(\beta)} \cdot \int_0^t x^{\beta-1}e^{-x}dx$.

Su función de riesgo es

$$h(t) = \frac{\lambda^\beta t^{\beta-1}e^{-\lambda t}}{\Gamma(\beta)\big(1-\Gamma^*(\lambda t,\beta)\big)},$$

la cual para $\beta > 1$ es monótona creciente, $h(0) = 0$ y $\lim_{t\to\infty} h(t) = \lambda$, mientras que para $\beta < 1$ es monótona decreciente, $\lim_{t\to 0} h(t) = \infty$ y $\lim_{t\to\infty} h(t) = \lambda$.

Estimación no paramétrica

La estimación no paramétrica, por otra parte, consiste en dar, para cada uno de los tiempos de vida observados (no censurados), el valor estimado de la **tasa de supervivencia**, esto es, el número de supervivientes entre el total de individuos.

Veamos un ejemplo sencillo de estimación de la curva de supervivencia en una base de datos pequeña, en la que se proporciona el tiempo de seguimiento (en meses) de cada paciente tras la aparición de una enfermedad, teniendo en cuenta que ningún dato es censurado y los tiempos de seguimiento, por tanto, se corresponden con los verdaderos tiempos de vida tras dicha aparición.

Así, los datos serían los siguientes (t_i), ordenados de manera ascendente para facilitar los cálculos, junto con la estimación de la tasa de supervivencia ($\hat{S}(t_i)$):

t_i	13	17	20	20	24	32	36	45
$\hat{S}(t_i)$	7/8	6/8	4/8		3/8	2/8	1/8	0

La estimación de la curva de supervivencia sería, por tanto, la representada en la siguiente gráfica:

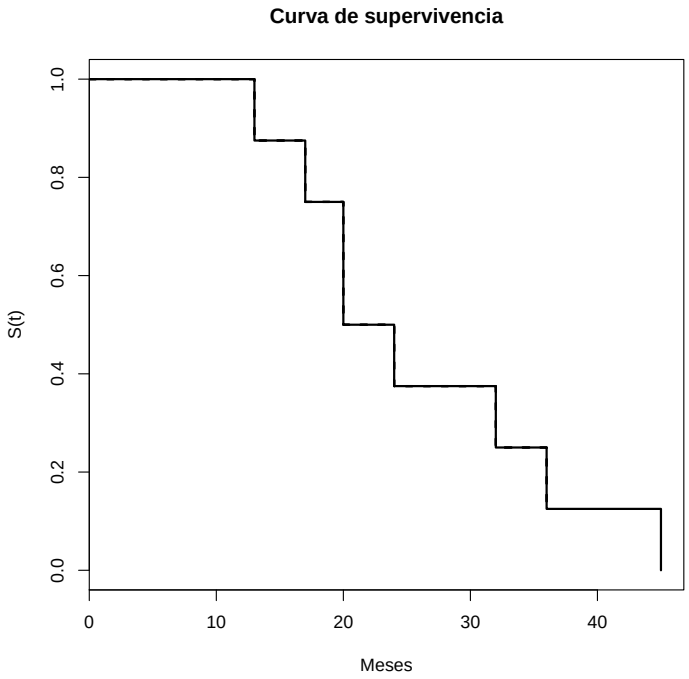


Figura 1.3: Estimación de la curva de supervivencia para los datos del ejemplo

1.2.3. Datos censurados

No siempre es posible tener el tiempo de vida completo de un paciente, ya que usualmente ocurre un suceso previo, llamado **censura**, que se presenta antes del suceso de interés. Éste puede deberse a varios motivos, como puede ser el fin del estudio (ya que un estudio no dura indefinidamente, por lo que no siempre da tiempo a que todos los pacientes experimenten el suceso de interés), un fallo diferente al que interesa (lo cual se denomina *presencia de riesgos competitivos*) o una pérdida de seguimiento como consecuencia de diversas posibles causas.

En estos casos, el tiempo de vida se observa, pero parcialmente, ya que sólamente se sabe que el tiempo de vida completo de este paciente sería superior al tiempo observado. Es decir, en los datos censurados lo que tenemos es simplemente una cota inferior del verdadero (y desconocido) tiempo de vida.

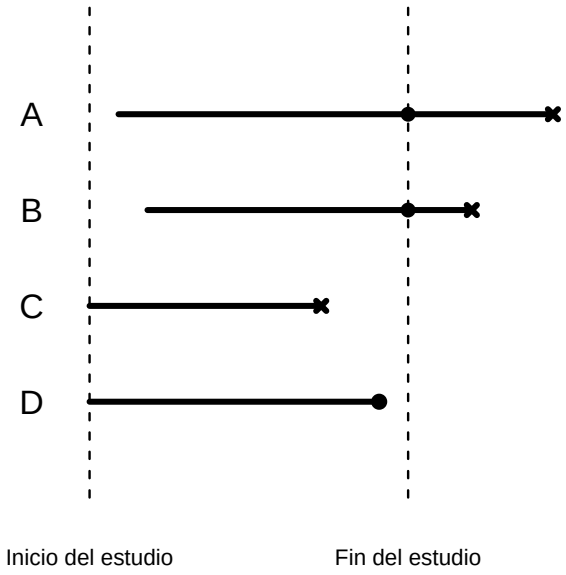


Figura 1.4: Ejemplo del seguimiento de cuatro pacientes (los círculos son datos censurados y las cruces señalan el tiempo en el que ocurrió el evento)

Hay diversos tipos de censura: censura por la derecha, censura por la izquierda y censura por intervalo. La que nos concierne en este trabajo es la primera, por lo que detallaremos a continuación en qué consiste esta censura.

En la **censura por la derecha**, el evento es observado solamente si éste ocurre antes de un tiempo predeterminado (llamado *fin del estudio*), que viene fijado al comienzo de éste, de manera independiente (*censura de tipo I*) o no (*censura de tipo II*) del tamaño de la muestra. Un ejemplo de este tipo de censura ocurre cuando se aplica un tratamiento a determinados pacientes y se desea conocer cuál es el tiempo transcurrido hasta la muerte de éstos. Debido al tiempo disponible y a los costes que el estudio supone, el investigador decide terminar el estudio a los 5 años del comienzo (sería pues censura de tipo I). Los tiempos de supervivencia registrados para los pacientes que fallecieron durante el periodo de estudio son los tiempos desde la aplicación del tratamiento hasta su muerte (individuo C); éstas son llamadas **observaciones exactas** o **no censuradas**. Los tiempos de supervivencia de los pacientes que continúan vivos al final del estudio no son conocidos

exactamente, pero son registrados al menos con la longitud del estudio (individuos A y B); éstas son llamadas **observaciones censuradas**. Algunos pacientes podrían perderse o morir accidentalmente (individuo D); en este caso se denominan también **observaciones censuradas**.

En el caso de censura por la derecha, es frecuente utilizar la siguiente notación: para un individuo específico i estudiado, se supone que tiene un tiempo de vida t_i y un tiempo de censura c_i. Se define entonces el tiempo observado y se denota por z_i al mínimo entre estos dos tiempos, esto es $z_i = \text{mín}\{t_i, c_i\}$. Así, el tiempo de vida exacto del individuo es conocido si y sólo si $t_i \leq c_i$, mientras que si $t_i > c_i$, sabemos que el sujeto es un superviviente y su tiempo de vida es censurado al final del estudio.

Los datos del estudio pueden estar representados por la pareja de variables $\{Z, \delta\}$, donde Z es la variable aleatoria considerada en el párrafo anterior y $\delta = I(Z = T)$, es decir, toma el valor 1 si el tiempo de vida observado es el real (*dato no censurado*) y 0 en caso contrario, esto es, que se haya producido una censura antes de la ocurrencia del evento ($Z = C$, *dato censurado*).

1.2.4. El estimador de Kaplan-Meier

La pregunta que cabe hacerse llegados a este punto es: *¿Cómo podemos estimar la función de supervivencia?* Pues bien, hay dos alternativas para ello: la alternativa paramétrica, por la cual podemos optar si conocemos a qué familia de distribuciones se ajusta nuestra función de supervivencia (entre las más conocidas se encuentran los modelos mencionados anteriormente), ajustando los parámetros correspondientes mediante estimadores máximo verosímiles; o la alternativa no paramétrica, entre las que destacan el estimador de Kaplan-Meier y el estimador de Fleming-Harrington.

Dado que el método de estimación de la curva ROC tiempo-dependiente propuesto como objetivo último de este trabajo requiere de la utilización de la estimación de Kaplan-Meier, a continuación daremos una nociones básicas acerca de ésta.

Si estamos ante una muestra en la que ninguna de las observaciones está censurada,

lo cual es muy poco común en la práctica, la función de supervivencia puede ser estimada por la función de supervivencia empírica, dada por

$$\hat{S}(t) = \frac{\text{N}^o \text{ total de sujetos que sobreviven más allá del tiempo } t}{\text{N}^o \text{ total de sujetos}} ,$$

o equivalentemente

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

donde $\hat{F}(t)$ es la función de distribución empírica, esto es

$$\hat{F}(t) = \frac{\text{N}^o \text{ total de individuos que han fallado antes del tiempo } t}{\text{N}^o \text{ total de individuos}} .$$

Un ejemplo de estimación de la función de supervivencia para datos no censurados es el propuesto en la *Figura 1.3*. Éste es un caso particular de la estimación de Kaplan-Meier cuando el conjunto de datos no posee censuras.

El estimador de Kaplan y Meier (1958) [13] es el estimador de la función de supervivencia más utilizado y se define, para el caso en el que la muestra contenga datos censurados por la derecha, como

$$\hat{S}_{KM}(t) = \prod_{i\,:\,t_i \leq t} \left(\frac{r(t_i) - d(t_i)}{r(t_i)} \right) = \prod_{i\,:\,t_i \leq t} \left(1 - \frac{d(t_i)}{r(t_i)} \right) ,$$

donde $r(t_i)$ es el número de individuos en riesgo y $d(t_i)$ es el número de muertes u ocurrencias del evento de interés correspondiente en el momento t_i. Cabe destacar que éste es el estimador máximo verosímil de $S(t)$.

Veamos a continuación un ejemplo sencillo de estimación de la curva de supervivencia en una base de datos pequeña que posee sujetos censurados. Los datos son los recogidos en la siguiente tabla, donde z_i denota el tiempo de seguimiento (en meses) de cada paciente tras la aparición de una enfermedad, ordenados de forma ascendente para facilitar los cálculos. Además, el signo $^+$ es la forma habitual de indicar que se trata de un dato censurado, por lo que en esos casos $z_i = c_i$, mientras que en el resto los z_i se corresponden con los tiempos reales de supervivencia t_i.

z_i	13	14^+	15^+	16	20	24^+	28	34
d_i	1	0	0	1	1	0	1	1
r_i	8	7	6	5	4	3	2	1
d_i/r_i	1/8	0	0	1/5	1/4	0	1/2	1
$1 - d_i/r_i$	7/8	1	1	4/5	3/4	1	1/2	0
$\hat{S}(t_i)$	7/8	7/8	7/8	7/10	21/40	21/40	21/80	0

La estimación de la curva de supervivencia sería, por tanto, la representada en la siguiente gráfica, teniendo en cuenta que los puntos "+" indican los datos censurados:

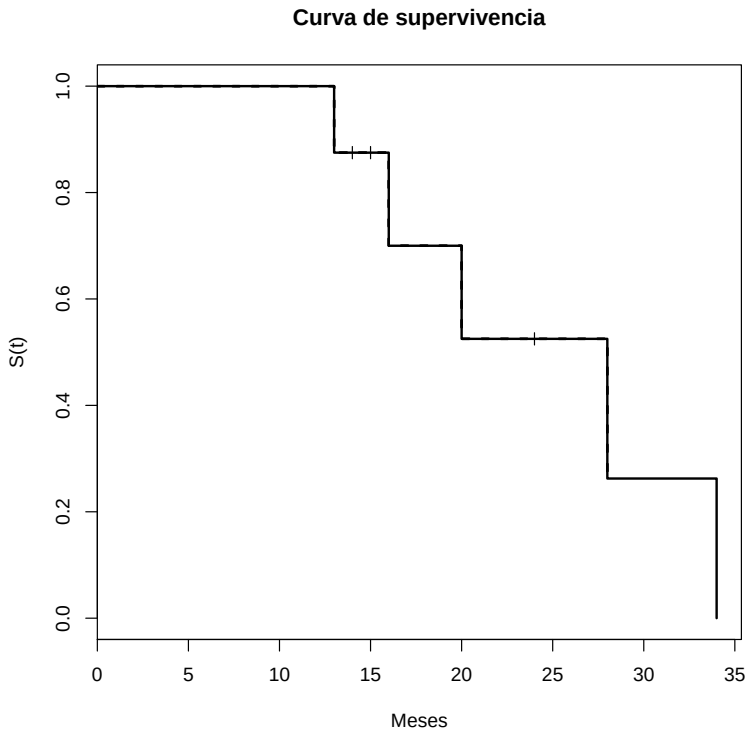


Figura 1.5: Estimación de la curva de supervivencia para los datos del ejemplo

Volviendo al marco teórico, componiendo el estimador de Kaplan-Meier con la función logaritmo,

$$log(\hat{S}(t)) = \sum_{i\,:\,t_i \leq t}^{k} log\left(1 - \frac{d(t_i)}{r(t_i)}\right) .$$

Puesto que la probabilidad de ocurrencia del evento es independiente entre los distintos intervalos de tiempo, la varianza de esta expresión viene dada por

$$Var\left(log(\hat{S}(t))\right) = \sum_{i\,:\,t_i \leq t}^{k} Var\left(log\left(1 - \frac{d(t_i)}{r(t_i)}\right)\right) .$$

A través de la fórmula de Greenwood se concluye que la varianza del estimador de Kaplan-Meier es igual a

$$Var\left(\hat{S}_{KM}(t)\right) = \hat{S}_{KM}(t) \sum_{i\,:\,t_i \leq t} \frac{d(t_i)}{r(t_i)(r(t_i) - d(t_i))} ,$$

por lo que un intervalo de confianza al $1 - \alpha\,\%$ para $S(t)$, para un valor concreto de t, está dado por

$$\left(\hat{S}_{KM}(t) - z_{1-\alpha/2} EE\left(\hat{S}_{KM}(t)\right), \hat{S}_{KM}(t) + z_{1-\alpha/2} EE\left(\hat{S}_{KM}(t)\right)\right)$$

donde EE denota el error estándar, esto es, $\sqrt{Var/n}$, siendo Var la varianza calculada anteriormente.

Se muestran a continuación las dos curvas de supervivencia estimadas anteriormente junto con sus intervalos de confianza al 95 %:

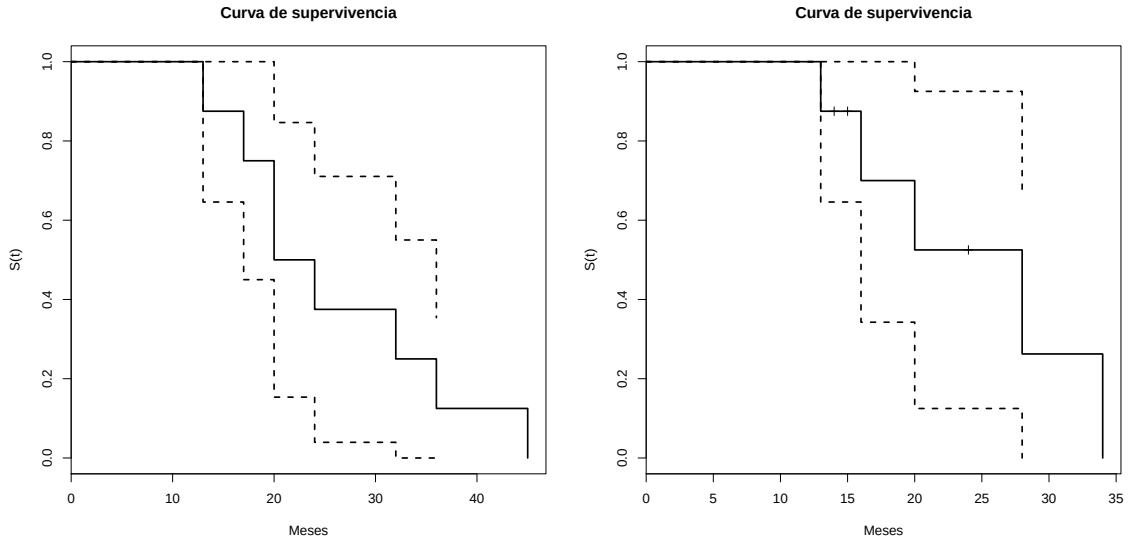


Figura 1.6: Estimación de las curvas de supervivencia para los datos de los ejemplos anteriores junto con sus intervalos de confianza

Cabe citar también que una parte importante del análisis de supervivencia es la creación de gráficos de curvas de supervivencia para cada grupo de interés. Sin embargo, la comparación de curvas de supervivencia entre dos grupos debe estar basada en tests estadísticos no paramétricos, entre los que destaca el test *logrank*, y no basarse únicamente en impresiones gráficas.

1.2.5. Regresión de Cox o modelo de riesgos proporcionales

El **modelo de Cox** es una técnica estadística cuyo objetivo principal es estudiar la relación entre la supervivencia de los pacientes y varias variables explicativas. Cuando se utiliza este modelo para analizar la supervivencia de los individuos en un ensayo clínico, éste nos permite separar los efectos del tratamiento de los efectos de otras variables. El modelo puede utilizarse también si se conoce de antemano que hay otras variables, aparte del tratamiento, que están influyendo en la supervivencia y que no pueden ser controladas fácilmente en el ensayo clínico. Por tanto, utilizando este modelo se puede mejorar la estimación del efecto del tratamiento. Los tiempos de supervivencia no están dirigidos tanto al desarrollo de un síntoma particular o la reincidencia de una enfermedad, sino al tiempo hasta la muerte u ocurrencia de un determinado evento.

El test *logrank* que mencionábamos en el apartado anterior no puede utilizarse para explorar y ajustar por los efectos de varias variables, como pueden ser la edad o la duración de la enfermedad, las cuales es sabido que suelen afectar en gran medida a la supervivencia. El ajuste por otras variables que se conoce que afectan al tiempo de supervivencia puede mejorar la precisión con la cual estamos estimando el efecto del tratamiento, si es el caso.

El método de regresión introducido por Cox se utiliza para estudiar varias variables a la vez. Es conocido también como **modelo de riesgos proporcionales**. En resumidas cuentas, este procedimiento modela los tiempos de supervivencia, o siendo más específicos, la función de riesgos, según las variables explicativas.

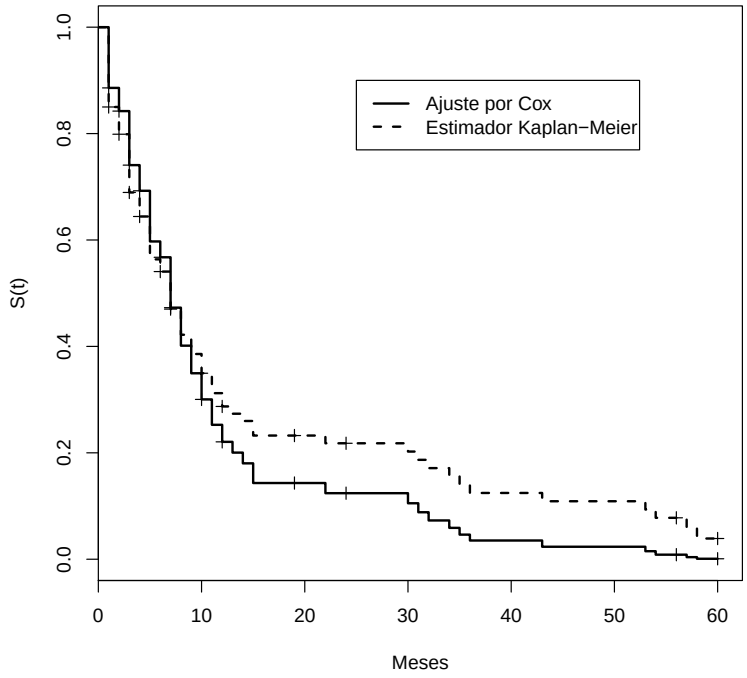


Figura 1.7: Ejemplo de estimación de la curva de supervivencia a través del estimador de Kaplan-Meier y del modelo de Cox, ajustando por dos covariables (edad, consumo de drogas)

Recordemos que la función de riesgo $h(t)$, definida en las subsecciones anteriores, es la probabilidad de que un individuo experimente un evento, como puede ser la muerte, en un intervalo de tiempo muy pequeño, dado que el individuo ha sobrevivido al menos hasta el comienzo de ese intervalo. Puede interpretarse asimismo como el riesgo de morir en el instante t.

El modelo de regresión de Cox es similar al modelo de regresión múltiple (en el que recordemos que se pretende describir la relación entre los valores de una variable, llamada dependiente, y los de otras, llamadas independientes), excepto porque ahora la variable dependiente es la función de riesgo en un tiempo dado.

Si tenemos varias variables explicativas (vector y), podemos expresar el riesgo de morir en el instante t como

$$h(t) = h_0(t) \cdot e^{\beta t y}$$

donde $h_0(t)$ se denomina **función de riesgo basal** y β es el vector de parámetros de la regresión. La función de riesgo basal se corresponde con la probabilidad de muerte u ocurrencia del evento cuando todas las variables explicativas son cero. Los parámetros de la regresión, por su parte, explican el cambio proporcional esperado en el riesgo, en relación a los cambios en las variables explicativas. Estos parámetros suelen estimarse utilizando el método de máxima verosimilitud.

Tomando logaritmos, el modelo de regresión de Cox se puede expresar como

$$log(h(t)) = log(h_0(t)) + \beta^t y \ .$$

La hipótesis de que la relación entre la variable dependiente $(h(t))$ y las variables explicativas (y) es constante se denomina *riesgos proporcionales* (en inglés *proportional hazards*). Esto significa que las funciones de riesgo de dos individuos en cualquier instante de tiempo son proporcionales; en otras palabras, que si un individuo tiene un riesgo doble de muerte en un tiempo inicial que otro individuo, se sabe que a lo largo todos los tiempos venideros, el riesgo de muerte del primero seguirá siendo el doble que el del segundo. Esta hipótesis de riesgos proporcionales debe ser probada mediante tests estadísticos. Lo más sencillo es representar gráficamente los estimadores de Kaplan-Meier de las curvas en la misma gráfica, y ver si se cruzan o no. Si se cruzan es porque no se cumple la hipótesis inicial de riesgos proporcionales. Sin embargo, este modelo gráfico es muy subjetivo, y puede venir muy determinado por el tamaño muestral, por lo que para mejorar un poco esta representación, se suele representar el logaritmo del opuesto del logaritmo de la estimación de la función de supervivencia $(log(-log(\hat{S}_{KM}(t))))$ frente al logaritmo del tiempo de supervivencia $(log(t))$, que proporcionará curvas paralelas si los riesgos son proporcionales en los grupos.

Era sabido, por lo visto en secciones anteriores. que el conocer la función de supervivencia y la función de riesgos o función de riesgos acumulada es equivalente, gracias a la siguiente relación:

$$H(t) = -log(S(t)) \quad \Rightarrow \quad S(t) = e^{-H(t)} \ .$$

En el caso del modelo de Cox, conocemos la expresión de la función de riesgo, por lo

tanto

$$H(t) = \int_0^t h(u)du = \int_0^t h_0(u)e^{\beta^t y} = e^{\beta^t y}\int_0^t h_0(u)du = H_0(t)e^{\beta^t y}$$

$$\Rightarrow \quad S(t \mid y) = e^{-H_0(t)e^{\beta^t y}} \ .$$

Una propiedad importante de este modelo es la siguiente:

Al hacer el cociente entre $h(t \mid Y_i = 0)$ y $h(t \mid Y_i = 1)$, considerando que el resto de componentes de Y son iguales a cero, se obtiene que

$$\frac{h(t \mid Y_i = 1)}{h(t \mid Y_i = 0)} = \frac{h_0(t)e^{\beta_i}}{h_0(t)} = e^{\beta_i} \ ,$$

por lo que el vector e^{β} indica cuánto varía la función de riesgo en un mismo tiempo t al hacer variar en una unidad alguna de las covariables, cuando el resto toman el valor cero.

El modelo de regresión de Cox se considera un procedimiento semi-paramétrico, ya que la función de riesgo basal $h_0(t)$ no tiene por qué estar especificada. Debido a este hecho, se utiliza un parámetro diferente para cada tiempo de supervivencia t. Como la función de riesgo no está restringida a una forma específica, el modelo semi-paramétrico admite gran flexibilidad y, por esta razón, es muy utilizado. Sin embargo, si se asume que los datos siguen una determinada distribución de probabilidad, las inferencias basadas en esta suposición serán mucho más precisas, esto es, los estimadores de las funciones de riesgo tendrán errores estándar mucho más pequeños y, por tanto, unos intervalos de confianza más estrechos.

Un modelo de riesgos proporcionales paramétrico se basa en las mismas suposiciones que el modelo de Cox, pero asumiendo, además, que la función de riesgo basal, $h_0(t)$, puede ser parametrizada de acuerdo a un determinado modelo de distribución de los tiempos de supervivencia. Algunas de las distribuciones de la función de supervivencia más utilizadas en estos casos son: la distribución exponencial, la de Weibull y la de Gompertz, ya que éstas cumplen la propiedad de riesgos proporcionales.

Capítulo 2

Curva ROC tiempo-dependiente

Cuando la variable que estamos midiendo es binaria, en nuestro caso D (que denota la verdadera prevalencia o no de una enfermedad, el fallecimiento o no de un individuo, etc.), es decir, el diagnóstico verdadero, la exactitud de un biomarcador X se suele resumir a través de las proporciones de buena clasificación, definidas como sensibilidad, $P(X > x \mid D = 1)$, y especificidad, $P(X \leq x \mid D = 0)$, asociadas a cada punto de corte x del biomarcador. Es este punto de corte el que marca un criterio para clasificar a los sujetos como positivos ($X > x$) o negativos ($X \leq x$). Cuando el valor de x no está indicado a priori, puede caracterizarse el espectro completo de sensibilidades y especificidades a través de la curva ROC, la cual, como hemos indicado en el capítulo anterior, representa la *proporción de verdaderos positivos* (*sensibilidad*) frente a la *proporción de falsos positivos* (*1 - especificidad*) para todo $x \in \mathbb{R}$.

En este capítulo veremos, en primer lugar, las propuestas existentes para la generalización de los conceptos de sensibilidad y especificidad cuando aplicamos éstas a tiempos de supervivencia. Se darán definiciones de sensibilidad y especificidad en términos del verdadero tiempo de supervivencia T, tratando los datos censurados de una forma adecuada para conseguir estimaciones válidas. Posteriormente mostraremos que una elección determinada de las definiciones de verdaderos positivos y falsos positivos en función del tiempo, da lugar a los diferentes modelos de curvas ROC tiempo-dependientes.

2.1. Extensiones de la sensibilidad y la especificidad

Existen varias extensiones de la sensibilidad y la especificidad para datos de supervivencia. Ahora, en vez de tener una salida binaria $D = 0$ ó $D = 1$, el tiempo de supervivencia puede interpretarse como una salida binaria que varía con el tiempo, esto es, la pertenencia de los individuos a uno u otro grupo varía en función de tiempo. Por ejemplo, alguien que está vivo en un estudio realizado durante un año, puede no pertenecer al grupo control si se realiza el mismo estudio considerando cinco años. La notación utilizada a continuación es la siguiente: t_i denota el tiempo real de supervivencia del individuo i-ésimo, independientemente de si en la muestra ha sido o no censurado con anterioridad. Las extensiones mostradas a continuación son clasificadas de acuerdo a cómo se definen los casos y los controles en un instante t. Así,

- **Casos incidentes** ($\mathbb{I}$): un sujeto es considerado *caso incidente* para el instante t cuando su tiempo real de ocurrencia del evento es exactamente ese instante, esto es, $t_i = t$.

- **Casos acumulativos** ($\mathbb{A}$): un sujeto es considerado *caso acumulativo* para el instante t cuando la ocurrencia del evento es previa a este instante, o exactamente éste, es decir, $t_i \leq t$.

- **Controles estáticos** ($\mathbb{E}$): un sujeto es considerado *control estático* cuando su tiempo real de supervivencia es mayor que un tiempo t_0 fijado previamente, esto es, $t_i > t_0$. Suele considerarse t_0 un valor fijo alto, en cuyo caso la interpretación de estos individuos sería como supervivientes a largo plazo. Cabe notar que esta definición es independiente del instante t, por lo que independientemente del t que estemos considerando, los *controles estáticos* serán siempre los mismos.

- **Controles dinámicos** ($\mathbb{D}$): un sujeto es considerado *control dinámico* cuando su tiempo de supervivencia es mayor que el tiempo t, esto es, $t_i > t$. Estos individuos no presentan el evento antes del tiempo t, aunque podrían presentarlo más adelante, en cuyo caso el individuo pasaría a ser considerado un *caso acumulativo*.

Teniendo en cuenta estas nociones, las definiciones de sensibilidad y especificidad en cada uno de estos casos serían las siguientes:

- **Sensibilidad incidente** ($\mathbb{I}$): $S_E^{\mathbb{I}}(x_0, t) = P(X > x_0 \,|\, T = t)$.
- **Sensibilidad acumulativa** ($\mathbb{A}$): $S_E^{\mathbb{A}}(x_0, t) = P(X > x_0 \,|\, T \leq t)$.
- **Especificidad estática** ($\mathbb{E}$): $S_P^{\mathbb{E}}(x_0) = P(X \leq x_0 \,|\, T > t_0)$.
- **Especificidad dinámica** ($\mathbb{D}$): $S_P^{\mathbb{D}}(x_0, t) = P(X \leq x_0 \,|\, T > t)$.

En este trabajo se considerará únicamente el caso en el que tenemos un único valor del biomarcador X para cada individuo, pero en el enfoque *incidente/dinámico* veremos que estos conceptos se pueden generalizar para el caso longitudinal, en el que se tiene el valor del biomarcador en distintos instantes de tiempo, esto es, $X(t)$.

2.1.1. Incidente/estática

Etzioni et al. (1999) [4] y *Slate y Turnbull (2000)* [29] adoptaron conceptos alternativos de la sensibilidad y la especificidad tiempo-dependientes usando las siguientes definiciones:

$$S_E^{\mathbb{I}}(x_0, t) = P(X > x_0 \,|\, T = t) \ ,$$

$$S_P^{\mathbb{E}}(x_0) = P(X \leq x_0 \,|\, T > t_0) \ .$$

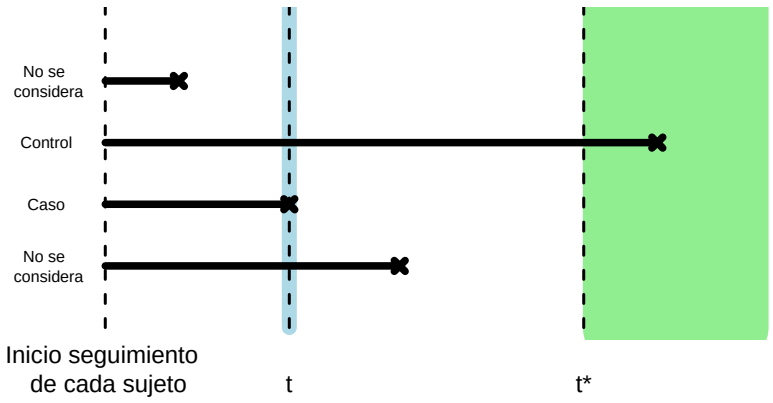


Figura 2.1: Clasificación de casos y controles en el enfoque incidente/estático

En base a estas expresiones, los controles son siempre los mismos, pero los casos están estrechamente ligados con el instante t considerado, es más, los individuos son clasificados como casos sólamente para un único instante t concreto. Los casos se clasifican conforme el tiempo en el que ocurre el evento, mientras que los controles están definidos como aquellos sujetos en los que no ocurre el evento durante un periodo de seguimiento fijado $(0, t_0)$.

Estas definiciones facilitan el uso de los métodos de regresión estándar para caracterizar la sensibilidad y la especificidad, ya que el tiempo del evento T puede utilizarse como una covariable. Para estimar los cuantiles de la distribución condicional del marcador X, dado un tiempo de ocurrencia del suceso $T = t$, los autores arriba referenciados consideraron la utilización de métodos paramétricos asumiendo una distribución normal, pero permitiendo que la media y la varianza fuesen funciones del tiempo medido, del estado real de la enfermedad y del tiempo de evento para los casos.

Las curvas ROC de este tipo se definen como

$$\mathcal{R}_t^{\mathbb{I}/\mathbb{E}}(p) = S_E^{\mathbb{I}}\big(\left[1 - S_P^{\mathbb{E}}\right]^{-1}(p, t_0), t\big) \qquad (0 \leq p \leq 1)$$

donde $\left[1 - S_P^{\mathbb{E}}\right]^{-1}(p, t_0) = \text{ínf}\{x : \left[1 - S_P^{\mathbb{E}}\right](x, t_0) \leq p\}$.

La expresión anterior define la curva ROC incidente/estática como una función de p, donde p es la probabilidad de falsos positivos, esto es, la probabilidad de que un control (sujeto que no presenta el evento hasta pasado el tiempo fijo t_0) sea clasificado como caso. Para calcular el valor de esta función ($\mathcal{R}_t$) en cada punto p, lo que se hace es hallar el ínfimo de los puntos de corte del biomarcador que hacen que la probabilidad de falsos positivos sea a lo sumo p, y posteriormente calcula la sensibilidad incidente para ese punto de corte en el instante t, es decir, la probabilidad de que un caso (sujeto al que le ocurre el evento en el preciso instante t) sea correctamente clasificado como tal.

En el año 2006, *Cai et al.* [3] propuso métodos para estimar la sensibilidad y la especificidad tiempo-dependientes cuando el tiempo de seguimiento está censurado. Más recientemente, *Zheng y Heagerty* [10] han propuesto métodos de regresión de cuantiles,

los cuales minimizan las hipótesis de distribuciones paramétricas de los métodos anteriores.

2.1.2. Incidente/dinámica

Según este enfoque, los conceptos de sensibilidad y especificidad tiempo-dependientes vienen dados por las siguientes definiciones:

$$S_E^{\mathbb{I}}(x_0, t) = P(X > x_0 \,|\, T = t) \ ,$$

$$S_P^{\mathbb{D}}(x_0, t) = P(X \leq x_0 \,|\, T > t) \ .$$

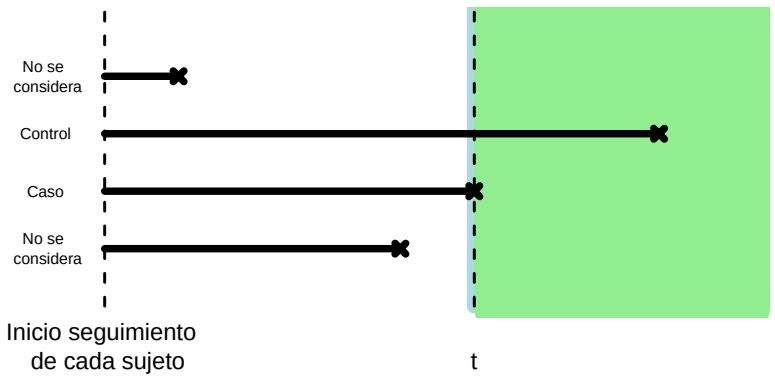


Figura 2.2: Clasificación de casos y controles en el enfoque incidente/dinámico

En este enfoque, según el tiempo t considerado va variando de forma creciente, un sujeto que en un primer momento (t pequeño) es un control, pasa a jugar el papel de caso cuando el instante t considerado coincide con el instante de ocurrencia del evento en este sujeto, es decir, $t_i = t$, y en cuanto t es mayor que t_i, ese individuo ya no se considera en el cálculo de la sensibilidad ni la especificidad.

Aquí, la sensibilidad mide la proporción esperada de sujetos con un valor del biomarcador X mayor que x_0, entre la subpoblación de individuos que fallecen en el instante t, mientras que la especificidad mide la proporción de sujetos con un valor del biomarcador X menor o igual que x_0, entre aquellos que sobreviven más allá del tiempo t.

Las curvas ROC de este tipo se definen como

$$\mathcal{R}_t^{\mathbb{I}/\mathbb{D}}(p) = S_E^{\mathbb{I}}\Big(\left[1 - S_P^{\mathbb{D}}\right]^{-1}(p,t), t\Big) \qquad (0 \leq p \leq 1)$$

donde $\left[1 - S_P^{\mathbb{D}}\right]^{-1}(p,t) = \text{ínf}\{x_0 : \left[1 - S_P^{\mathbb{D}}\right](x_0,t) \leq p\}$.

Hay una característica particular de esta alternativa que cabe destacar, y es que la sensibilidad incidente y la especificidad dinámica están basadas en el conjunto de sujetos que están en riesgo en el tiempo t. Además, estas definiciones permiten la extensión a covariables tiempo-dependientes, definiendo la sensibilidad como $S_E^{\mathbb{I}}(x_0,t) = P(X(t) > x_0 \,|\, T = t)$ y la especificidad como $S_P^{\mathbb{D}}(x_0,t) = P(X(t) \leq x_0 \,|\, T > t)$, donde $X(t)$ es un biomarcador longitudinal.

2.1.3. Acumulativa/dinámica

En el año 2000, *Heagerty et al.* [11] propusieron versiones de la sensibilidad y la especificidad tiempo-dependientes usando las siguientes definiciones:

$$S_E^{\mathbb{A}}(x_0,t) = P(X > x_0 \,|\, T \leq t)\ ,$$

$$S_P^{\mathbb{D}}(x_0,t) = P(X \leq x_0 \,|\, T > t)\ .$$

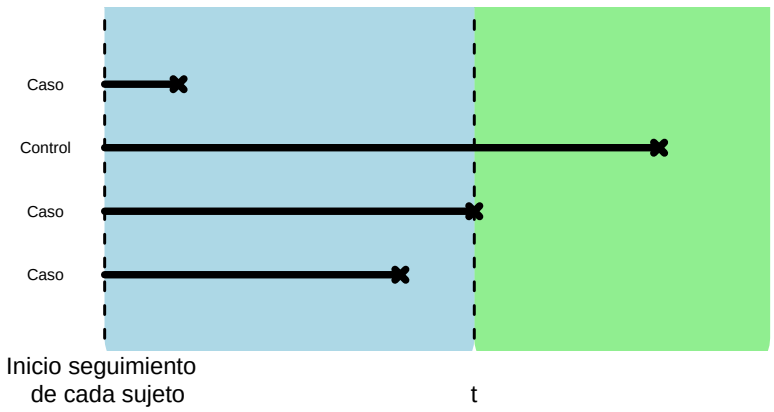


Figura 2.3: Clasificación de casos y controles en el enfoque acumulativo/dinámico

Utilizando este enfoque, para cualquier tiempo fijo t, toda la población será clasificada como caso o como control en base al estado de vida en ese instante. Además, a medida que hacemos variar t, cada individuo toma el papel de control para los tiempos tempranos, mientras t es menor que el tiempo de ocurrencia del evento ($t < t_i$), y seguidamente, en cuanto $t \geq t_i$, pasa a ser considerado como caso.

El enfoque acumulativo/dinámico es apropiado cuando el interés principal del investigador es discriminar entre los sujetos que mueren antes o en el instante t, y los que sobreviven más allá de t.

Las curvas ROC de este tipo se definen como

$$\mathcal{R}_t^{\mathbb{A}/\mathbb{D}}(p) = S_E^{\mathbb{A}}\big(\left[1 - S_P^{\mathbb{D}}\right]^{-1}(p,t), t\big) \qquad (0 \leq p \leq 1)$$

donde $\left[1 - S_P^{\mathbb{D}}\right]^{-1}(p,t) = \text{ínf}\{x_0 : \left[1 - S_P^{\mathbb{D}}\right](x_0,t) \leq p\}$.

Si no hay datos censurados, $\mathcal{R}_t^{\mathbb{A}/\mathbb{D}}(p)$ puede estimarse usando la función de distribución empírica del marcador $\hat{F}_X$ en cada uno de los grupos caso y control.

En caso de existencia de datos censurados en la base de datos, las propuestas existentes, junto con una novedosa, son presentadas en el siguiente capítulo.

Capítulo 3

Estimación de la curva ROC acumulativa/dinámica

En el enfoque acumulativo/dinámico de la curva ROC se utilizarán todos los sujetos en cualquier tiempo fijo t, lo cual no ocurría con los otros dos enfoques, como se puede observar en la *Figura 2.1* y *Figura 2.2* anteriores (a los cuales nos hemos referido como *sujetos no especificados*).

Cuando la información es completa, esto es, no existen datos censurados en nuestra base de datos, se pueden definir directamente los estimadores empíricos de la sensibilidad acumulativa y la especificidad dinámica como sigue

$$\hat{S}_E^{\mathbb{A}}(x_0, t) = \frac{\#\{x_i > x_0 \ \wedge \ t_i \leq t\}}{\#\{t_i \leq t\}} \, ,$$

$$\hat{S}_P^{\mathbb{D}}(x_0, t) = \frac{\#\{x_i \leq x_0 \ \wedge \ t_i > t\}}{\#\{t_i > t\}}$$

donde # denota el cardinal del conjunto.

El problema principal para la estimación de la curva ROC acumulativa/dinámica $\mathcal{R}_t^{\mathbb{A}/\mathbb{D}}$ es la presencia de datos censurados. Los individuos a los que les ocurre el evento antes de t, y los que tienen un tiempo de seguimiento mayor que t, se clasifican directamente en el grupo positivo (casos) y en el grupo negativo (controles), respectivamente. Sin embargo, no es tan evidente cómo debemos tratar a los individuos censurados antes de t, como es el caso del sujeto D en la siguiente figura.

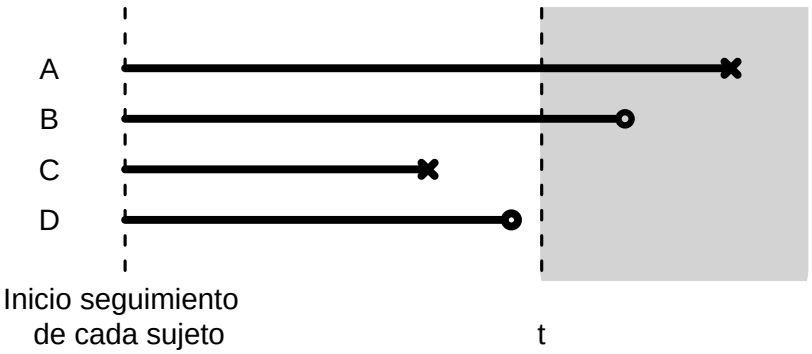


Figura 3.1: Situación esquemática: los círculos representan datos censurados y las cruces indican el tiempo en el que ocurrió el evento

Quizás, en este caso, la primer tentación es definir un estimador basado en la tradicional función de supervivencia de Kaplan-Meier. Ésta es la propuesta que viene detallada en la siguiente sección.

Antes de dar comienzo a la explicación de la siguiente serie de propuestas que han ido surgiendo a lo largo del tiempo para resolver esta situación, se formaliza la notación a utilizar:

Partimos de un muestra aleatoria simple de tamaño N compuesta por sujetos que pueden ser descritos a través de la siguiente terna:

$$y_i = \{z_i, \delta_i, x_i\} \qquad \forall i = 1, ..., N$$

donde

- z_i denota el tiempo observado, es decir, $z_i = \text{mín}\{t_i, c_i\}$, siendo t_i el tiempo de ocurrencia del evento y c_i el tiempo de censura;
- δ_i es el estado, que toma el valor 1 si $z_i = t_i$ (dato no censurado) y 0 si $z_i = c_i$ (dato censurado); y
- x_i denota el valor del biomarcador.

3.1. Primer intento: Basado en el estimador de Kaplan-Meier

Podemos utilizar el teorema de Bayes para reescribir la sensibilidad y la especificidad como

$$S_E^{\mathbb{A}}(x_0,t) = P(X > x_0 \,|\, T \leq t) = \frac{[1 - S(t \mid X > x_0)]P(X > x_0)}{1 - S(t)} \,,$$

$$S_P^{\mathbb{D}}(x_0,t) = P(X \leq x_0 \,|\, T > t) = \frac{S(t \mid X \leq x_0)P(X \leq x_0)}{S(t)}$$

donde $S(t)$ es la función de supervivencia, $S(t) = P(T > t)$, y $S(t \mid X > x_0)$ es la función de supervivencia condicionada a que $X > x_0$.

Un estimador no paramétrico de la función $S(t)$, enormemente conocido, es el dado por Kaplan y Meier en 1958 (definido en el capítulo 1). Recordemos que el estimador de Kaplan-Meier hace uso de toda la información de los datos, incluyendo las observaciones censuradas, para estimar la función de supervivencia.

Un estimador sencillo para la sensibilidad y la especificidad en el tiempo t es el dado por la combinación del estimador de Kaplan-Meier y la función de distribución empírica del biomarcador X, esto es

$$\hat{S}_E^{\mathbb{A}}(x_0,t) = \hat{P}_{KM}(X > x_0 \,|\, T \leq t) = \frac{[1 - \hat{S}_{KM}(t \mid X > x_0)] \cdot [1 - \hat{F}_X(x_0)]}{1 - \hat{S}_{KM}(t)} \,,$$

$$\hat{S}_P^{\mathbb{D}}(x_0,t) = \hat{P}_{KM}(X \leq x_0 \,|\, T > t) = \frac{\hat{S}_{KM}(t \mid X \leq x_0) \cdot \hat{F}_X(x_0)}{\hat{S}_{KM}(t)}$$

siendo $\hat{F}_X(x_0) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} I(X \leq x_0)$.

Uno de los problemas de este estimador es que no garantiza que la sensibilidad y la especificidad sean monótonas. Por definición, tenemos que si $x_0' > x_0$ entonces $P(X > x_0 \mid T \leq t) \geq P(X > x_0' \mid T \leq t)$. Sin embargo, los estimadores resultantes del teorema de Bayes y el estimador de Kaplan-Meier podrían violar esta monotonía, ya que el estimador de la probabilidad $P(X > x_0 \ \wedge \ T > t)$ dado por $\hat{S}_{KM}(t \mid X > x_0) \cdot [1 - \hat{F}_X(x_0)]$ puede no dar lugar a una distribución bivariada válida.

Un ejemplo en el que se puede ver este hecho es el siguiente:

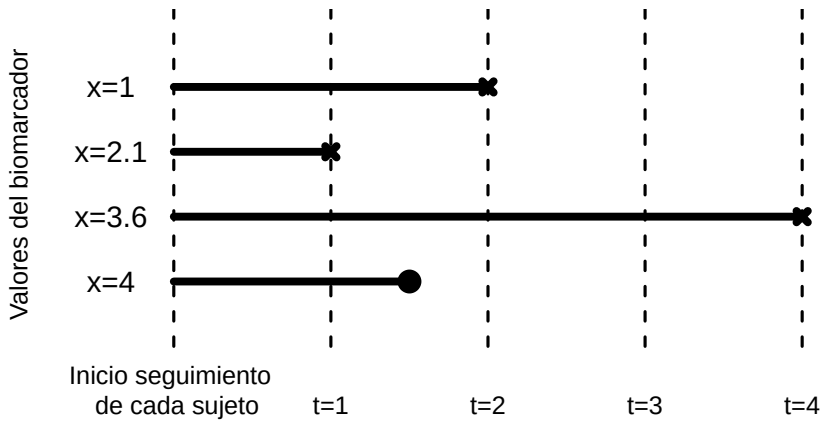


Figura 3.2: Situación esquemática: el círculo representa el tiempo de censura y las cruces indican el tiempo en el que ocurrió el evento

Vamos a calcular los estimadores $\hat{S}_{KM}(t \mid X > 0)$ y $\hat{S}_{KM}(t \mid X > 1)$ siguiendo el esquema de la tabla correspondiente a la *Figura 1.5*:

$\hat{S}_{KM}(t \mid X > 0)$

z_i	1	1.5^+	2	4
d_i	1	0	1	1
r_i	4	3	2	1
d_i/r_i	1/4	0	1/2	1
$1 - d_i/r_i$	3/8	1	1/2	0
$\hat{S}(t_i)$	3/4	3/4	3/8	0

$\hat{S}_{KM}(t \mid X > 1)$

z_i	1	1.5^+	4
d_i	1	0	1
r_i	3	2	1
d_i/r_i	1/3	0	1
$1 - d_i/r_i$	2/3	1	0
$\hat{S}(t_i)$	2/3	2/3	0

Así,

$$\hat{P}(X > 0 \ \wedge \ T > 3) = \hat{S}_{KM}(3 \mid X > 0) \cdot [1 - \hat{F}_X(0)] = \frac{3}{8} \cdot (1 - 0) = \frac{3}{8} \, ,$$

$$\hat{P}(X > 1 \ \wedge \ T > 3) = \hat{S}_{KM}(3 \mid X > 1) \cdot [1 - \hat{F}_X(1)] = \frac{2}{3} \cdot \left(1 - \frac{1}{4}\right) = \frac{1}{2} \, .$$

Por tanto, se tiene que la probabilidad estimada siguiente es negativa, lo cual es una incongruencia

$$\hat{P}(X \in (0,1] \ \wedge \ T > 3) = \hat{P}(X > 0 \ \wedge \ T > 3) - \hat{P}(X > 1 \ \wedge \ T > 3) = \frac{3}{8} - \frac{1}{2} = -\frac{1}{8} \, .$$

Este problema puede ser atribuido al hecho de que, dado que el conjunto por el que se condiciona, $X > x_0$, cambia, la redistribución a la derecha de la función masa de probabilidad asociada a las observaciones censuradas también varía.

En este caso, el estimador de Kaplan-Meier distribuye la probabilidad de la observación censurada ($x_4 = 4$ con $t_4 > 1.5$) entre aquellos individuos cuyos tiempos de observación son mayores que el tiempo de censura c_4. Cuando se tiene en cuenta el subconjunto $X > 0$, esta probabilidad se distribuye entre $t = 2$ y $t = 4$, mientras que si se considera el subconjunto $X > 1$, ésta se asigna completamente al tiempo $t = 4$.

Este cambio en la redistribución puede provocar inconsistencias que dan lugar a estimaciones de probabilidades negativas, produciendo por tanto curvas ROC no monótonas.

El segundo gran problema que presenta este método de estimación de la curva ROC acumulativa/dinámica es que el estimador de Kaplan-Meier condicionado, $\hat{S}_{KM}(t \mid X > x_0)$, asume que el proceso de censura es independiente del biomarcador X. Esta hipótesis puede no cumplirse en la práctica, como ocurre cuando la intensidad de las medidas de seguimiento se ven influenciadas por las medidas del biomarcador.

3.2. Segundo intento: Basado en el estimador de vecinos próximos (KNN)

Otra propuesta para el estimador de $\mathcal{R}_t^{\mathbb{A}/\mathbb{D}}$ es el basado en el estimador de la función de distribución bivariada $F(x_0, t) = P(X \leq x_0 \ \wedge \ T \leq t)$, o equivalentemente de $S(x_0, t) = P(X > x_0 \ \wedge \ T > t)$, propuesto por *Akritas* en el año 1994 [1].

Este estimador se basa en la representación de $S(x_0, t)$ de la siguiente forma

$$S(x_0, t) = \int_{x_0}^{\infty} S(t \mid X = u) \, dF_X(u)$$

donde F_X denota la función de distribución del biomarcador.

Como muestra *Akritas* en su artículo, un estimador de la función anterior vendría dado por

$$\hat{S}_{\lambda_N}(x_0, t) = \frac{1}{N} \sum_{i=1}^{N} \hat{S}_{\lambda_N}(t \mid X = x_i) \cdot I(x_i > x_0)$$

donde $\hat{S}_{\lambda_N}(x_0, t)$ es un estimador adecuado de la función de supervivencia condicionada, caracterizado por un parámetro λ_N.

Con esta propuesta, se define el estimador ponderado de Kaplan-Meier como

$$\hat{S}_{\lambda_N}(t \mid X = x_i) = \sum_{k\,:\,t_k \leq t} \left(1 - \frac{\sum_{j=1}^{N} K_{\lambda_N}(x_j, x_i) \cdot I(z_j = t_k) \cdot \delta_j}{\sum_{j=1}^{N} K_{\lambda_N}(x_j, x_i) \cdot I(z_j \leq t_k)} \right)$$

donde $K_{\lambda_N}(X_j, X_i)$ es una función núcleo que depende de un parámetro de suavizado λ_N. En su artículo, *Akritas*[1] propone la siguiente función núcleo

$$K_{\lambda_N}(X_j, X_i) = I\left(-\lambda_N < \hat{F}_X(X_i) - \hat{F}_X(X_j) < \lambda_N \right)$$

donde $2\lambda_N \in (0, 1)$ representa el porcentaje de observaciones que son incluidas en cada entorno.

Hay otras opciones para la función núcleo, sin embargo utilizando los vecinos más próximos se tiene que las estimaciones de la curva ROC resultantes son invariantes frente a transformaciones monótonas del biomarcador X. Utilizando como K_{λ_N} el núcleo del vecino más próximo, el citado autor presenta cotas para el parámetro λ_N, suficientes para alcanzar la convergencia débil del estimador de la función de distribución bivariada. Cabe destacar que el estimador del vecino más próximo (NNE) es un estimador semiparamétrico eficiente.

Los estimadores de la sensibilidad y la especificidad resultantes son los siguientes

$$\hat{S}_E^{\mathbb{A}}(x_0, t) = \hat{P}_{\lambda_N}(X > x_0 \mid T \leq t) = \frac{\left(1 - \hat{F}_X(x_0)\right) - \hat{S}_{\lambda_N}(x_0, t)}{1 - \hat{S}_{\lambda_N}(t)},$$

$$\hat{S}_P^{\mathbb{D}}(x_0, t) = \hat{P}_{\lambda_N}(X \leq x_0 \mid T > t) = 1 - \frac{\hat{S}_{\lambda_N}(x_0, t)}{\hat{S}_{\lambda_N}(t)}$$

donde $\hat{S}_{\lambda_N}(t) = \hat{S}_{\lambda_N}(x_0, t)$ con $x_0 = -\infty$.

Cabe notar que el numerador de $\hat{S}_E^{\mathbb{A}}(x_0, t)$ es igual a $\frac{1}{N}\sum_{i=1}^{N} I(X_i > x_0)\cdot(1 - S_{\lambda_N}(t \mid X = X_i))$, el cual es una función monótona creciente respecto de x_0, en comparación a lo que sucedía con el estimador basado en la estimación de Kaplan-Meier visto en la subsección anterior.

Otra ventaja importante del estimador NNE es que no supone que el proceso de censura es independiente de los valores del biomarcador X, al contrario nuevamente de lo que sucedía con el estimador de la propuesta anterior. Esta característica es muy relevante, sobre todo en la práctica, ya que estas variables suelen estar relacionadas.

Sin embargo, esta propuesta posee una gran desventaja: la elección del parámetro de suavizado del estimador, λ_N, queda en manos del investigador, por lo que no hay objetividad.

3.3. Tercer intento: Basado en el estimador de Nelson-Aalen

Más recientemente, *Wolf*, *Schmidt* y *Ulm* [37] propusieron un estimador basado en la curva de incidencia acumulativa (CIC) de Aalen.

Al igual que en la propuesta basada en el estimador de Kaplan-Meier, en ésta se parte también de la utilización del teorema de Bayes para representar la sensibilidad y la especificidad como sigue:

$$S_E^{\mathbb{A}}(x_0, t) = P(X > x_0 \mid T \leq t) = \frac{P(T \leq t \mid X > x_0) \cdot P(X > x_0)}{P(T \leq t)} ,$$

$$S_P^{\mathbb{D}}(x_0, t) = P(X \leq x_0 \mid T > t) = \frac{P(T > t \mid X \leq x_0) \cdot P(X \leq x_0)}{P(T > t)} .$$

Como se ha visto en la primera propuesta, no se pueden estimar estas probabilidades utilizando el estimador de Kaplan-Meier para toda la muestra. Para observaciones censuradas no se cumple que

$$P(T \leq t) = P(T \leq t \mid X \leq x_0) \cdot P(X \leq x_0) + P(T \leq t \mid X > x_0) \cdot P(X > x_0)$$

con proporciones fijas $P(X \leq x_0)$ (respectivamente $P(X > x_0)$), es decir, no verifican el teorema de la probabilidad total.

Para obtener la curva ROC en el tiempo t podemos estimar el número de eventos esperados antes de t para cada subconjunto $X \leq x_0$ y $X > x_0$, que denotaremos $e_0(x_0, t)$ y $e_1(x_0, t)$, respectivamente, de la siguiente forma

- Si $X \leq x_0 : e_0(x_0, t) = P(T \leq t \mid X \leq x_0) \cdot N_0(x_0)$,
- Si $X > x_0 : e_1(x_0, t) = P(T \leq t \mid X > x_0) \cdot N_1(x_0)$,

donde $N_0(x_0)$ y $N_1(x_0)$ son los tamaños muestrales de los conjuntos $\{X \leq x_0\}$ y $\{X > x_0\}$, respectivamente.

La sensibilidad y la especificidad podrían estimarse, por tanto, a través de las siguientes expresiones:

$$\hat{S}_E^{\mathbb{A}}(x_0, t) = \frac{e_1(x_0, t)}{e_0(x_0, t) + e_1(x_0, t)} ,$$

$$\hat{S}_P^{\mathbb{D}}(x_0, t) = \frac{N_0(x_0) - e_0(x_0, t)}{N - e_0(x_0, t) - e_1(x_0, t)} .$$

La otra posibilidad es estimar el número esperado de eventos en el tiempo t usando todas las observaciones de eventos hasta el tiempo t y calculando el número adicional de eventos que se tendrían teniendo en cuenta aquellos que fueron censurados antes de t.

Para observaciones censuradas, se tiene que estimar la probabilidad de ocurrencia del evento entre el tiempo de censura c y el tiempo t ($t > c$). Esta probabilidad $P(T \leq t \mid T > c)$ puede estimarse utilizando la función de riesgo $h(t)$ entre c y t.

Si ocurre la censura entre t_{k-1} y t_k, la probabilidad de ocurrencia del evento antes del tiempo t viene dada por

$$P(T \leq t \mid T > c) = P(c < T \leq t_k \mid T > c) + P(t_k < T \leq t \mid T > c) .$$

Para estimar el segundo sumando se utiliza la siguiente expresión

$$\frac{P(T > t_k) - P(T > t)}{P(T > c)} = \frac{S(t_k) - S(t)}{S(c)} ,$$

mientras que para estimar el primer sumando, $P(c < T \leq t_k \mid T > c)$ hay varias posibilidades. El cálculo exacto sería utilizando la definición de un proceso de Poisson $P(c < T \leq t_k \mid T > c) \approx \lambda_{k-1}(t_k - c)$ donde λ_{k-1} es la tasa del proceso.

Haciendo uso de esta definición, un sujeto censurado poco después de t_{k-1} tiene una probabilidad más alta de ocurrencia del evento que un individuo cuyo tiempo de censura es inmediatamente anterior a t_k. Para facilitar los cálculos, asumimos que el tiempo de censura es el punto medio del intervalo (t_{k-1}, t_k), obteniendo por tanto que una posible estimación de la probabilidad $P(c < T \leq t_k \mid T > c)$ sería $\frac{t_k - t_{k-1}}{2}\lambda_{k-1}$.

Bajo la hipótesis de que la probabilidad anterior es la misma si la censura se produce al inicio o al final del intervalo $(t_{k-1}, t_k]$, ésta se reduce a la siguiente expresión:

$$P(T \leq t \mid T > c) = P(c < T \leq t_k \mid T > c) + P(t_k < T \leq t \mid T > c)$$

$$= \frac{S(c) - S(t_k)}{S(c)} + \frac{S(t_k) - S(t)}{S(c)} = 1 - \frac{S(t)}{S(c)} = 1 - \frac{e^{-H(t)}}{e^{-H(c)}} = 1 - e^{-(H(t) - H(c))} .$$

Utilizando la aproximación de Taylor de primer orden, se tiene que bajo la hipótesis de que $H(t) - H(c) << 1$, la probabilidad anterior se puede aproximar del siguiente modo

$$P(T \leq t \mid T > c) \approx H(t) - H(c) .$$

Para calcular esta diferencia, en vez de utilizar el estimador de Kaplan-Meier para estimar el valor de la función de riesgo acumulada en un instante t, $H(t)$, se puede recurrir al estimador de Nelson-Aalen, que es el siguiente

$$H(t) = \sum_{i\,:\,t_i \leq t} h(t_i) \sum_{i\,:\,t_i \leq t} h_i = \sum_{i\,:\,t_i \leq t} \frac{d(t_i)}{r(t_i)}$$

donde $d(t_i)$ es el número de ocurrencias del evento en el momento t_i y $r(t_i)$ es el número de individuos en riesgo en este mismo momento.

Este cálculo debe hacerse para todas las observaciones censuradas antes de t y teniendo en cuenta que se hará en cada grupo por separado, esto es, considerando los sujetos con $X > x_0$ por un lado, y los sujetos con $X \leq x_0$ por otro.

Teniendo en cuenta estas probabilidades y el número de eventos observados, se puede calcular el número esperado de eventos antes del tiempo t. A partir de todos estos valores, se puede obtener la sensibilidad y la especificidad para los distintos valores de x_0, y así construir la curva ROC para un tiempo t determinado, como sigue

$$\hat{S}_E^{\mathbb{A}}(x_0, t) = \frac{E_1(x_0, t)}{E_0(x_0, t) + E_1(x_0, t)} ,$$

$$\hat{S}_P^{\mathbb{D}}(x_0, t) = \frac{N_0(x_0) - E_0(x_0, t)}{N - E_0(x_0, t) - E_1(x_0, t)} ,$$

donde $E_0(x_0, t) = \sum_{i\,:\,t_i \leq t} \big(d_0(x_0, t_i) + e_0(x_0, t_i)\big)$ y $E_1(x_0, t) = \sum_{i\,:\,t_i \leq t} \big(d_1(x_0, t_i) + e_1(x_0, t_i)\big)$, siendo $e_0(x_0, t)$ y $e_1(x_0, t)$ los valores definidos anteriormente, es decir, el número de eventos observados antes de t para los conjuntos $\{X \leq x_0\}$ y $\{X > x_0\}$, respectivamente; y $d_0(x_0, t)$ y $d_1(x_0, t)$ el número de eventos esperados antes de t para los conjuntos $\{X \leq x_0\}$ y $\{X > x_0\}$, respectivamente.

El principal problema de esta propuesta es el mismo que uno de los que poseía la basada en el estimador de Kaplan-Meier, y es que la sensibilidad y la especificidad no siempre resultan funciones monótonas respecto a x_0. Sin embargo, un método propuesto por los autores para solventar esta situación es por medio de la regresión isotónica (*Salanti y Ulm, 2005* [28]). Si falla la monotonía, se puede aplicar el algoritmo $PAVA$ (del inglés *Pooling Adjacent Violator Algorithm*), definiendo los pares de puntos contiguos $(\hat{S}_E^{\mathbb{A}}(x_0, t), \hat{S}_P^{\mathbb{D}}(x_0, t))$ y $(\hat{S}_E^{\mathbb{A}}(x_0 + 1, t), \hat{S}_P^{\mathbb{D}}(x_0 + 1, t))$ de la siguiente manera:

$$(\hat{S}_E^{\mathbb{A}}(x_0, t), \hat{S}_P^{\mathbb{D}}(x_0, t)) = (\hat{S}_E^{\mathbb{A}}(x_0 + 1, t), \hat{S}_P^{\mathbb{D}}(x_0 + 1, t))$$

$$= \left(\frac{S_E^{\mathbb{A}}(x_0, t) + S_E^{\mathbb{A}}(x_0 + 1, t)}{2}, \frac{S_P^{\mathbb{D}}(x_0, t) + S_P^{\mathbb{D}}(x_0 + 1, t)}{2} \right) ,$$

resolviendo así este problema.

3.4. Nueva propuesta

Esta nueva propuesta pretende acabar con todos los problemas que presentaba la primera. Para ello, se tendrá en cuenta toda la muestra, incluyendo los sujetos censurados, pero tratando a éstos de una forma diferente: en vez de asignarlos de forma íntegra a uno de los grupos (caso/control) para un determinado tiempo t, se considerará la probabilidad de que pertenezcan a cada uno de los grupos. Es para esta probabilidad para la cual se dan dos posibles estimadores.

Partiendo nuevamente del teorema de Bayes, podemos reescribir la especificidad y la sensibilidad de la curva acumulativa/dinámica del siguiente modo

$$S_E^{\mathbb{A}}(x_0,t) = \frac{P(X > x_0 \ \wedge \ T \leq t)}{P(T \leq t)} = \frac{\int_0^1 P(X > x_0 \ \wedge \ T \leq t \mid y)\, dF_Y}{\int_0^1 P(T \leq t \mid y)\, dF_Y}$$

$$S_P^{\mathbb{D}}(x_0,t) = \frac{P(X \leq x_0 \ \wedge \ T > t)}{P(T > t)} = \frac{\int_0^1 P(X \leq x_0 \ \wedge \ T > t \mid y)\, dF_Y}{\int_0^1 P(T > t \mid y)\, dF_Y}$$

donde y es el valor observado ($\{z, \delta, x\}$) y F_Y es su función de distribución.

Una posible estimación del contenido de la integral del numerador de las expresiones anteriores de sensibilidad y especificidad, para una muestra dada, es la siguiente

$$P(X > x_0 \ \wedge \ T \leq t \mid y_i) = P(T \leq t \mid y_i) \cdot I_{(x_0,\infty)}(x_i)$$

$$= \big(1 - P(T > t \mid y_i)\big) \cdot I_{(x_0,\infty)}(x_i) \qquad \forall i = 1, ..., N$$

$$P(X \leq x_0 \ \wedge \ T > t \mid y_i) = P(T > t \mid y_i) \cdot I_{(-\infty,x_0]}(x_i) \qquad \forall i = 1, ..., N$$

siendo $I_A(x)$ la función indicador, que vale 1 si $x \in A$ y 0 en caso contrario.

Por tanto, si denotamos por $\hat{P}_i$ (en verdad sería $\hat{P}_i(N)$, ya que dependerá del tamaño muestral) el estimador de la probabilidad $P(T > t \mid y_i)$, los estimadores empíricos de la sensibilidad y la especificidad pueden expresarse como

$$\hat{S}_E^{\mathbb{A}}(x_0,t) = \frac{\sum_{i=1}^{N}(1-\hat{P}_i)\cdot I_{(x_0,\infty)}(x_i)}{\sum_{i=1}^{N}(1-\hat{P}_i)} ,$$

$$\hat{S}_P^{\mathbb{D}}(x_0,t) = \frac{\sum_{i=1}^{N}\hat{P}_i\cdot I_{(-\infty,x_0]}(x_i)}{\sum_{i=1}^{N}\hat{P}_i} .$$

Naturalmente, si $z_i > t$, esto es, el tiempo observado (ya sea el tiempo real de ocurrencia del evento o el tiempo de censura) es posterior a t, se tiene que $\hat{P}_i = 1$; y si $t_i < t$, es decir, el tiempo de ocurrencia del evento, para aquellos sujetos no censurados, se concluye que $\hat{P}_i = 0$.

Cuando la información es completa, es decir, no hay datos censurados, estos estimadores son los usuales. Además, la consistencia de los estimadores anteriores es directa si el estimador de $P(T > t \mid y_i)$ tiene buenas condiciones, particularme, si $|\hat{P}_i - P(T > t \mid y_i)|$ converge a 0 en probabilidad. Por otra parte, es obvio que la curva ROC resultante subsana los problemas comentados de la estimación basada en el estimador de Kaplan-Meier, es decir, para un tiempo determinado t, la curva es monótona y siempre toma valores menores o iguales que 1.

Esta propuesta proporciona dos procedimientos distintos para estimar la probabilidad anterior en aquellos individuos que no están completamente definidos, es decir, aquellos individuos censurados antes del tiempo t considerado. Las propuestas son las siguientes:

- Propuesta semiparamétrica, utilizando la regresión de Cox:

  Mediante el modelo de regresión de Cox de riesgos proporcionales, podemos estimar la función de riesgo como

  $$h(t) = h_0(t)\cdot e^{\beta X}$$

  donde X es el valor del biomarcador, esto es, se toma como única covariable en el modelo el valor de X. Utilizando esta estimación de la función de riesgo, calculamos $\hat{P}_i$ de la siguiente manera

  $$\hat{P}_i = \frac{\hat{S}(t \mid X = x_i)}{\hat{S}(z_i \mid X = x_i)}$$

  donde $\hat{S}$ es la función de supervivencia estimada por el modelo de regresión de Cox.

- Propuesta no paramétrica, usando directamente el estimador de Kaplan-Meier:

  Puesto que estamos asumiendo que el biomarcador X es una variable aleatoria continua, no se puede utilizar el estimador de Kaplan-Meier directamente para estimar la supervivencia de la expresión anterior, ya que está condicionada por el conjunto $X = x_i$. En este caso, se reemplaza por $X \leq x_i$, y lo que se propone es seleccionar los individuos que satisfagan $X \leq x_i$ y utilizar, ahora sí, el estimador de Kaplan-Meier para obtener $\hat{P}_i$ como sigue

  $$\hat{P}_i = \frac{\hat{S}_{KM}(t)}{\hat{S}_{KM}(z_i)}$$

  donde $\hat{S}_{KM}$ es la función de supervivencia estimada por Kaplan-Meier, referida únicamente a aquellos sujetos que cumplen $X \leq x_i$.

Los individuos censurados juegan un papel fundamental en esta propuesta. Cabe notar que la incertidumbre completa acerca de a qué grupo (caso/control) pertenece el sujeto i-ésimo en el tiempo t (la cual ocurre cuando $\hat{P}_i = 1/2$) implica, con independencia del grupo al cual sea asignado, un error de 1/2. Por tanto, en estos casos, la capacidad diagnóstica del biomarcador será limitada.

Este hecho aparece reflejado en la siguiente figura, para la cual se ha considerado la base de datos utilizada por *Wolf, Schmidt y Ulm (2001) [37]* con el mismo propósito. Esta base de libre acceso dentro del paquete `KMsurv` del software estadístico `R` contiene el tiempo hasta la muerte y la edad (utilizada como un biomarcador de mortalidad en este caso) de un total de 863 pacientes a los que se les ha transplantado un riñón.

En el gráfico de la izquierda se representan las siguientes estimaciones de la curva ROC acumulativa/dinámica para el tiempo t igual a 9 años:

- D_I: Curva ROC usual teniendo en cuenta solamente los datos no censurados, que constituyen un total de 157 pacientes, entre los cuales sólo 17 son controles (diagnóstico verdadero),

- K_M: Estimación basada en el método de Kaplan-Meier, y

- A_K: Estimación basada en el método del vecino más próximo (KNN).

Las diferencias observadas entre las estimaciones basadas en KNN y las basadas en el método de estimación de la función de riesgos acumulada propuesto por *Nelson-Aalen* son prácticamente despreciables cuando se toma como parámetro de suavizado del KNN, λ_N, uno muy próximo a cero.

En el gráfico de la derecha, por su parte, aparecen representadas:

- D_I: Nuevamente la curva ROC usual teniendo en cuenta solamente los datos no censurados,
- N_C: Estimación de $R_9^{\mathbb{A}/\mathbb{D}}$ mediante el nuevo método propuesto basado en la regresión de Cox, y
- N_K: Estimación de $R_9^{\mathbb{A}/\mathbb{D}}$ mediante el nuevo método propuesto basado en el estimador de Kaplan-Meier.

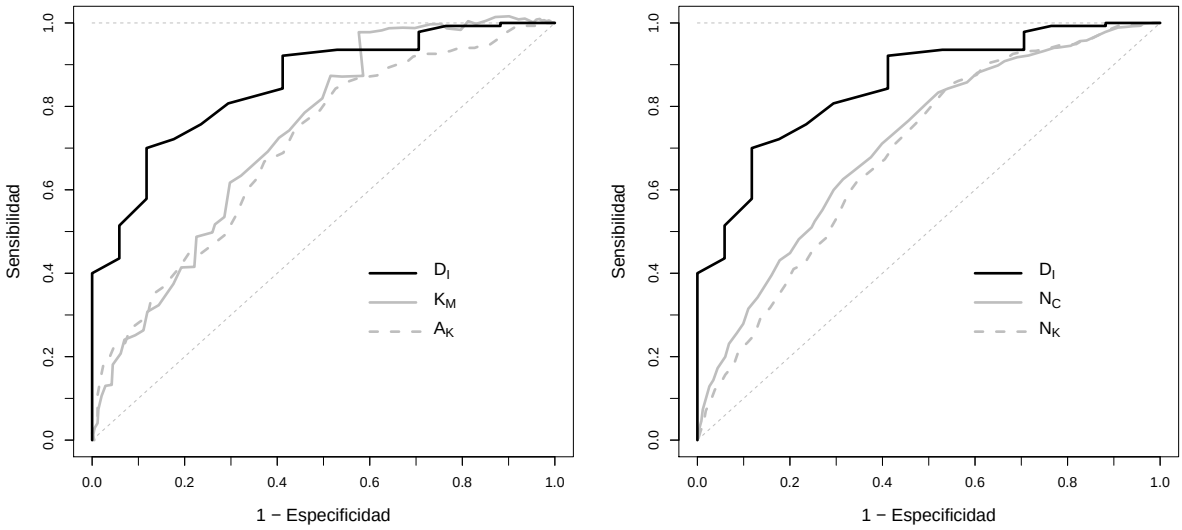


Figura 3.3: Varias estimaciones de la curva ROC: D_I = usual curva ROC sin los datos censurados; K_M = basado en el estimador de Kaplan-Meier; A_K = basado en el método KNN con $\lambda_N = 0.01 \cdot N^{-1/5}$; N_C = nueva propuesta basada en la regresión de Cox; N_K = nueva propuesta basada en el estimador de Kaplan-Meier

Capítulo 4

Estudio de simulación

Con el objetivo de estudiar el comportamiento práctico de la metodología propuesta, se ha llevado a cabo un estudio de simulación mediante el conocido método de Monte Carlo. De forma similar a lo realizado por *Heatherty y Zheng* en su artículo [10], y siguiendo con la notación que se ha venido utilizando hasta el momento, se consideran las siguientes distribuciones y parámetros:

- La distribución conjunta del tiempo de supervivencia real y el valor del biomarcador, $(log(T), X)$, se corresponde con la distribución Normal bivariada estándar, con coeficiente de correlación ρ. Esto es,

$$\begin{pmatrix} log(T) \\ X \end{pmatrix} \equiv \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) .$$

  Se han considerado dos casos para el coeficiente de correlación: $\rho = -1/4$ y $\rho = -3/4$. Ambos negativos, para indicar que están relacionados de forma inversa, es decir, tiempos de vida altos están asociados a valores pequeños del biomarcador, como suele ocurrir en la práctica (y así lo venimos considerando desde el primer capítulo). Además, por ser el segundo valor de ρ más próximo a 1 en valor absoluto, se tiene una mayor correlación entre ambas variables en ese segundo caso.

- Se han tomado dos tamaños muestrales diferentes: $N = 100$ y $N = 200$.

- La distribución del tiempo de censura, $log(C)$, se ha considerado también Normal con desviación típica $\sigma = 1$. En cuanto a la media, se han tomado dos valores distintos, dando lugar así a dos porcentajes de datos censurados diferentes en la muestra, teniendo en cuenta que se han considerado independientes el tiempo de censura y el tiempo real de supervivencia.

  - $\mu = 0$, así

$$P(C < T) = P(C - T < 0) = P(log(C) - log(T) < 0)$$
$$= P(\mathcal{N}(0,1) - \mathcal{N}(0,1) < 0) = P(\mathcal{N}(0, \sqrt{2}) < 0) = 0.5 \, .$$

    Es decir, la mitad de los datos son censurados antes de la ocurrencia del evento.

  - $\mu = 1.19$, así

$$P(C < T) = P(C - T < 0) = P(log(C) - log(T) < 0)$$
$$= P(\mathcal{N}(1.19,1) - \mathcal{N}(0,1) < 0) = P(\mathcal{N}(1.19, \sqrt{2}) < 0) = 0.2 \, .$$

    Es decir, un 20 % de los datos son censurados antes de la ocurrencia del evento.

- La relación entre el tiempo de censura y el valor del biomarcador se impone mediante dos valores de la covarianza entre $log(C)$ y X diferentes: $\tau = 0$ (no hay relación) y $\tau = 1/4$.

Se han realizado estimaciones de la curva ROC acumulativa/dinámica, $R_t^{\mathbb{A}/\mathbb{D}}$, para los siguientes tiempos: t tal que $log(t) = -1$ ($\Rightarrow t = 1/e$), $log(t) = 0$ ($\Rightarrow t = 1$) y $log(t) = 1$ ($\Rightarrow t = e$). Estas estimaciones se han llevado a cabo a través de los distintos métodos presentados en el capítulo anterior:

- D_I: Curva ROC usual teniendo en cuenta solamente los datos no censurados,
- K_M: Estimación basada en el método de Kaplan-Meier (primera sección),
- A_K: Estimación basada en el método del vecino más próximo (KNN) con parámetro de suavizado $\lambda_N = 0.1 \cdot N^{-1/5}$ (segunda sección),

- N_C: Estimación mediante el nuevo método propuesto basado en la regresión de Cox,
- N_K: Estimación mediante el nuevo método propuesto basado en el estimador de Kaplan-Meier.

En la siguiente figura se muestran las curvas ROC teóricas sin censuras de las cuales han sido extraídas las muestras.

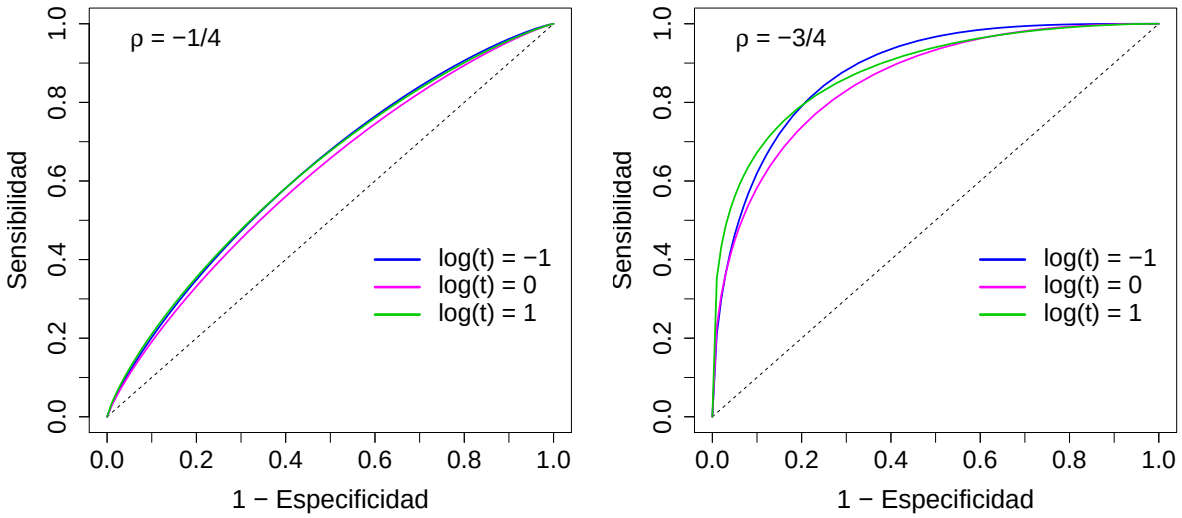


Figura 4.1: Curvas ROC reales $\mathcal{R}(p)$. En el gráfico de la izquierda, se ha considerado que el coeficiente de correlación ρ entre $log(T)$ y X es $-1/4$. En el gráfico de la izquierda, se ha considerado $\rho = -3/4$

Se observa que el biomarcador X en la gráfica situada a la derecha tiene mejor capacidad diagnóstica para clasificar correctamente a los sujetos en casos y controles, en cada uno de los tres tiempos t considerados. Esto es lógico, puesto que en el segundo caso, la correlación entre el tiempo de supervivencia y el valor del biomarcador es más estrecha ($\rho = -3/4$).

Además, la diferencias entre las curvas $\mathcal{R}^{\mathbb{A}/\mathbb{D}}_{t=1/e}(p)$, $\mathcal{R}^{\mathbb{A}/\mathbb{D}}_{t=1}(p)$ y $\mathcal{R}^{\mathbb{A}/\mathbb{D}}_{t=e}(p)$ son pequeñas en ambas gráficas, siendo aparentemente aún menores en la figura de la izquierda, lo cual vendría justificado por la misma razón que el párrafo anterior.

En las siguientes tablas se recogen los resultados obtenidos de la media y la desviación típica (expresados como *media* $\pm$ *desv. típ.*) de los cálculos de la siguiente integral a lo largo de las distintas muestras:

$$\sqrt{N}\int_0^1 |\hat{\mathcal{R}}(p) - \mathcal{R}(p)|\, dp$$

donde $\mathcal{R}$ es la curva ROC acumulativa/dinámica real y $\hat{\mathcal{R}}$ su estimación por medio de cada uno de los cinco métodos enumerados anteriormente, considerando las distintas combinaciones de los parámetros comentados.

La primera tabla (Tabla 4.1) se corresponde con los valores resultantes cuando se considera que los tiempos de censura y los valores del biomarcador son independientes entre sí, esto es, $\tau = 0$; mientras que la segunda tabla (Tabla 4.2) recoge los resultados cuando se considera que la covarianza entre el logaritmo de los tiempos de censura, $log(C)$, y los valores del biomarcador, X, es $\tau = 1/4$.

Para remarcar visualmente los resultados de la simulación, se han señalado en color azul y verde las mejores aproximaciones, consideradas como aquellas cuya media de la distancia entre $\mathcal{R}$ y $\hat{\mathcal{R}}$ en la métrica $\mathcal{L}_1$ es menor. Se observa que los cinco métodos estudiados se comportan de manera similar en la mayoría de los casos. Sin embargo, en casi todos ellos, las mejores estimaciones (en el sentido que estamos estudiando) se obtienen a partir del nuevo método propuesto y, además, entre ellos, parece que el basado en la regresión de Cox obtiene mejores resultados. Las mayores diferencias entre unos métodos y otros se obtienen, como cabía esperar, cuando los porcentajes de censura son altos (mitad de los datos censurados antes de t), sobre todo al estimar $\mathcal{R}_{log(t)=1}$. En estos casos, los mejores resultados provienen de las dos metodologías propuestas.

En la segunda tabla, en la cual se ha considerado que hay relación entre los tiempos de censura y los valores del biomarcador, los resultados son similares a los mostrados en la tabla anterior. Sin embargo, en algunos casos, la primera propuesta basada en el método de estimación de Kaplan-Meier (K_M) funciona mejor que la nueva obtenida a partir del mismo método de estimación (N_K). Por otra parte, cabe destacar que el método propuesto utilizando la regresión de Cox (N_C) siempre se encuentra entre los dos mejores.

N	ρ	%C	$log(t)$	K_M	A_K	D_I	N_C	N_K
100	-1/4	20 %	-1	0.834 ± 0.376	0.833 ± 0.399	0.827 ± 0.365	0.828 ± 0.373	0.830 ± 0.375
			0	0.613 ± 0.271	0.614 ± 0.289	0.625 ± 0.294	0.587 ± 0.266	0.591 ± 0.267
			1	0.934 ± 0.463	0.920 ± 0.460	0.887 ± 0.443	0.767 ± 0.392	0.767 ± 0.415
100	-3/4	20 %	-1	0.476 ± 0.214	0.474 ± 0.248	0.471 ± 0.214	0.468 ± 0.214	0.471 ± 0.214
			0	0.424 ± 0.182	0.429 ± 0.212	0.410 ± 0.179	0.390 ± 0.173	0.399 ± 0.181
			1	0.591 ± 0.320	0.623 ± 0.362	0.486 ± 0.283	0.411 ± 0.182	0.476 ± 0.176
100	-1/4	50 %	-1	0.868 ± 0.391	0.866 ± 0.420	0.853 ± 0.374	0.801 ± 0.368	0.835 ± 0.385
			0	0.744 ± 0.338	0.745 ± 0.358	0.750 ± 0.350	0.585 ± 0.299	0.614 ± 0.310
			1	1.519 ± 0.866	1.092 ± 0.542	1.123 ± 0.453	0.813 ± 0.483	0.956 ± 0.646
100	-3/4	50 %	-1	0.532 ± 0.231	0.492 ± 0.250	0.489 ± 0.216	0.448 ± 0.211	0.481 ± 0.221
			0	0.570 ± 0.243	0.533 ± 0.264	0.487 ± 0.215	0.365 ± 0.171	0.466 ± 0.243
			1	1.511 ± 1.065	1.051 ± 0.667	1.118 ± 1.009	0.391 ± 0.181	0.761 ± 0.490
200	-1/4	20 %	-1	0.821 ± 0.356	0.807 ± 0.381	0.831 ± 0.387	0.816 ± 0.354	0.818 ± 0.356
			0	0.618 ± 0.267	0.612 ± 0.286	0.630 ± 0.297	0.591 ± 0.260	0.595 ± 0.261
			1	0.950 ± 0.442	0.952 ± 0.472	0.997 ± 0.496	0.784 ± 0.371	0.784 ± 0.398
200	-3/4	20 %	-1	0.484 ± 0.201	0.499 ± 0.247	0.481 ± 0.201	0.478 ± 0.201	0.481 ± 0.201
			0	0.435 ± 0.181	0.435 ± 0.208	0.424 ± 0.177	0.400 ± 0.171	0.412 ± 0.179
			1	0.596 ± 0.258	0.618 ± 0.328	0.539 ± 0.245	0.428 ± 0.181	0.532 ± 0.285
200	-1/4	50 %	-1	0.866 ± 0.380	0.856 ± 0.407	0.879 ± 0.412	0.801 ± 0.358	0.835 ± 0.374
			0	0.749 ± 0.334	0.750 ± 0.359	0.784 ± 0.379	0.589 ± 0.300	0.621 ± 0.313
			1	1.599 ± 0.882	1.284 ± 0.638	1.453 ± 0.703	0.859 ± 0.509	0.995 ± 0.671
200	-3/4	50 %	-1	0.535 ± 0.224	0.525 ± 0.264	0.498 ± 0.209	0.456 ± 0.213	0.494 ± 0.219
			0	0.581 ± 0.238	0.529 ± 0.247	0.536 ± 0.223	0.375 ± 0.170	0.412 ± 0.263
			1	1.278 ± 0.755	1.177 ± 0.689	1.173 ± 0.642	0.457 ± 0.220	1.042 ± 0.517

Tabla 4.1: Media ± desviación típica de $0.01 \cdot \sqrt{N} \cdot \int_0^1 |\hat{\mathcal{R}}(p) - \mathcal{R}(p)|dp$, donde $\mathcal{R}$ es la curva ROC acumulativa/dinámica real y $\hat{\mathcal{R}}$ es su estimación, calculada a partir de 5000 iteraciones de Monte Carlo para $\tau = 0$. En color azul aparecen señaladas las mejores aproximaciones en cada caso, seguidas de las marcadas en verde.

N	ρ	%C	$log(t)$	K_M	A_K	D_I	N_C	N_K
100	-1/4	20 %	-1	0.828 ± 0.375	0.829 ± 0.402	0.829 ± 0.370	0.825 ± 0.375	0.830 ± 0.376
			0	0.601 ± 0.270	0.614 ± 0.294	0.614 ± 0.301	0.588 ± 0.269	0.604 ± 0.277
			1	0.867 ± 0.421	0.913 ± 0.446	0.914 ± 0.407	0.754 ± 0.384	0.823 ± 0.424
100	-3/4	20 %	-1	0.468 ± 0.207	0.465 ± 0.238	0.470 ± 0.207	0.466 ± 0.205	0.468 ± 0.205
			0	0.411 ± 0.179	0.421 ± 0.196	0.415 ± 0.182	0.389 ± 0.167	0.399 ± 0.169
			1	0.532 ± 0.291	0.536 ± 0.281	0.522 ± 0.334	0.398 ± 0.169	0.412 ± 0.208
100	-1/4	50 %	-1	0.830 ± 0.382	0.872 ± 0.420	0.845 ± 0.364	0.803 ± 0.372	0.855 ± 0.396
			0	0.657 ± 0.394	0.748 ± 0.354	0.741 ± 0.339	0.582 ± 0.299	0.672 ± 0.345
			1	1.253 ± 0.698	1.054 ± 0.495	1.078 ± 0.409	0.779 ± 0.462	1.123 ± 0.657
100	-3/4	50 %	-1	0.509 ± 0.235	0.484 ± 0.239	0.494 ± 0.225	0.450 ± 0.206	0.482 ± 0.205
			0	0.563 ± 0.257	0.499 ± 0.223	0.493 ± 0.235	0.368 ± 0.168	0.421 ± 0.189
			1	1.271 ± 0.909	0.698 ± 0.388	1.240 ± 1.041	0.343 ± 0.162	0.539 ± 0.369
200	-1/4	20 %	-1	0.822 ± 0.359	0.818 ± 0.384	0.832 ± 0.381	0.820 ± 0.359	0.824 ± 0.361
			0	0.605 ± 0.266	0.614 ± 0.289	0.638 ± 0.305	0.595 ± 0.266	0.612 ± 0.275
			1	0.870 ± 0.402	0.938 ± 0.452	1.029 ± 0.510	0.758 ± 0.374	0.836 ± 0.422
200	-3/4	20 %	-1	0.474 ± 0.203	0.493 ± 0.246	0.476 ± 0.204	0.472 ± 0.202	0.474 ± 0.201
			0	0.430 ± 0.188	0.430 ± 0.202	0.433 ± 0.189	0.406 ± 0.173	0.418 ± 0.176
			1	0.561 ± 0.264	0.547 ± 0.258	0.559 ± 0.288	0.423 ± 0.174	0.433 ± 0.202
200	-1/4	50 %	-1	0.831 ± 0.371	0.859 ± 0.405	0.881 ± 0.410	0.803 ± 0.364	0.855 ± 0.392
			0	0.674 ± 0.307	0.753 ± 0.364	0.822 ± 0.406	0.593 ± 0.305	0.692 ± 0.356
			1	1.294 ± 0.691	1.222 ± 0.570	1.360 ± 0.619	0.818 ± 0.493	1.217 ± 0.721
200	-3/4	50 %	-1	0.535 ± 0.230	0.497 ± 0.228	0.499 ± 0.206	0.451 ± 0.188	0.488 ± 0.192
			0	0.633 ± 0.274	0.515 ± 0.213	0.518 ± 0.221	0.385 ± 0.172	0.444 ± 0.193
			1	1.129 ± 0.678	0.786 ± 0.414	0.931 ± 0.797	0.391 ± 0.186	0.625 ± 0.442

Tabla 4.2: Media ± desviación típica de $0.01 \cdot \sqrt{N} \cdot \int_0^1 |\hat{\mathcal{R}}(p) - \mathcal{R}(p)|dp$, donde $\mathcal{R}$ es la curva ROC acumulativa/dinámica real y $\hat{\mathcal{R}}$ es su estimación, calculada a partir de 5000 iteraciones de Monte Carlo para $\tau = 1/4$. En color azul aparecen señaladas las mejores aproximaciones en cada caso, seguidas de las marcadas en verde.

Bajo estas líneas se muestra el código implementado en el software estadístico R para llevar a cabo las simulaciones.

```
library(survivalROC)
library(pROC)
library(mvtnorm)
library(survival)
library(ks)

##### COX

CX<- function(DT,t)
{
prob <- NULL
T <- DT[,1]
E <- DT[,2]
M <- DT[,3]
P <- which(T <= t & E==1)
N <- which(T > t)
IN <- which(T <= t & E==0)

if (length(IN)>0)
{
fit <- coxph(Surv(T, E) ~ M)
md <- survfit(fit, newdata=data.frame(DT))
prob <- 1:length(IN)
for (j in 1:length(IN))
{
f <- approxfun(c(min(md$time)-1,md$time),c(1,md$surv[,IN[j]]))
prob[j] <- f(t)/f(T[IN[j]])
if (is.na(prob[j])) prob[j]<- 1
}
}
cut <- sort(c(min(unique(sort(M)))-1,unique(sort(M)),max(unique(sort(M)))+1))
nS <- length(P) + sum(1-prob)
```

```
nE <- length(N) + sum(prob)
Se <- 1:length(cut)
Es <- 1:length(cut)
for (i in 1:length(cut))
{
Se[i] <- (sum(M[P]> cut[i]) + sum(1-prob[which(M[IN]>cut[i])]))/nS
Es[i] <- (sum(M[N]<= cut[i]) + sum(prob[which(M[IN]<=cut[i])]))/nE
}
list(TP=Se, TN=Es, prob=prob)
}

####  KAPLAN-MEIER

KM <- function(DT,t){
prob <- NULL
T <- DT[,1]
E <- DT[,2]
M <- DT[,3]
P <- which(T <= t & E==1)
N <- which(T > t)
F <- which(T <= t & E==0)

if (length(F)>0)
{
prob <- 1:length(F)
for (j in 1:length(F))
{
I<- which(M<=M[F[j]])
fit<- survfit(Surv(T[I],E[I])~1)
f <- approxfun(c(min(fit$time)-1,fit$time),c(1,fit$surv))
prob[j]<- f(t)/f(T[IN[j]])
}
if (is.na(prob[j])) prob[j]<- 1
}
cut <- sort(c(min(unique(sort(M)))-1,unique(sort(M)),max(unique(sort(M)))+1))
```

```
nS<- length(P) + sum(1-prob)
nE<- length(N) + sum(prob)
Se<- cut;
Es<- cut
for (i in 1:length(cut))
{
Se[i]<- (sum(M[P]> cut[i]) + sum(1-prob[which(M[F]>cut[i])]))/nS
Es[i]<- (sum(M[N]<= cut[i]) + sum(prob[which(M[F]<=cut[i])]))/nE
}
list(TP=Se, TN=Es, prob=prob)
}

##### METODO DIRECTO

rocDI<- function(DT,t)
{
I1<- which(DT[,1]<= t & DT[,2]==1)
I0<- which(DT[,1]>t)
if (length(I1)*length(I0)==0) {FP<- c(0,0.5,1); TP<- c(0,0.5,1)}
else {
M<- c(DT[I0,3],DT[I1,3])
R<- c(rep(0,length(I0)),rep(1,length(I1)))
r<- roc(R,M)
FP<- 1-r$spec; TP<- r$sens}
list( FP= FP, TP=TP)
}

##### PARAMETROS

B<- 5000; # Numero de iteraciones a considerar
Ekm<- matrix(0,nrow=B,ncol=3); Eak<- Ekm; Ep1<- Ekm; Edi<- Ekm; Ep2<- Ekm; Eco<- Ekm

c<- -0.75  # Coeficiente de correlacion entre log(T) y C ----> c <- -0.25 / c <- -0.75
ro<- 0.25 # Covarianza entre log(C) y X -----> ro <- 0 / ro <- 0.25
n<- 200; # Tamano muestral ------> n <- 100 / n <- 200
```

```
z<- seq(-5,5,0.1);
p<- seq(0,1,0.01);

##### CURVAS ROC REALES

mu<- c(0,0)
sigma<- matrix(c(1,c,c,1),ncol=2,nrow=2)
t<- -1
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),c(50,50),mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-50),c(50,t),mu,sigma)[1]/pnorm(t)
FR1<- approxfun(c(0,FP,1),c(0,TP,1))(p)
t<- 0
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),c(50,50),mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-50),c(50,t),mu,sigma)[1]/pnorm(t)
FR2<- approxfun(c(0,FP,1),c(0,TP,1))(p)
t<- 1
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),c(50,50),mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-50),c(50,t),mu,sigma)[1]/pnorm(t)
FR3<- approxfun(c(0,FP,1),c(0,TP,1))(p)

##### SIMULACIONES

for (b in 1:B)
{
mx<- matrix(c(1,c,0,c,1,ro,0,ro,1),ncol=3,nrow=3)
D<- rmvnorm(n,c(mu,0),mx) # El valor cero indica la media de la v.a. log(C) ------> 0 / 1.19

Z<- pmin(D[,1],D[,3]); # En cada fila compara ambos valores y toma el minimo
```

```
E<- (D[,1]<= D[,3])
DT<- cbind(Z,E,D[,2])

t<- -1
km<- survivalROC(Stime=exp(DT[,1]),status=DT[,2],marker = DT[,3],
                  predict.time = exp(t),method = "KM" )
FK<- approxfun(c(0,km$FP,1),c(0,km$TP,1))(p)
AK<- survivalROC(Stime=exp(DT[,1]),status=DT[,2],marker = DT[,3],
                  predict.time = exp(t),span = 0.1*n^(-0.20))
FA<- approxfun(c(0,AK$FP,1),c(0,AK$TP,1))(p)
DI<- rocDI(DT,t)
FD<- approxfun(c(0,DI$FP,1),c(0,DI$TP,1))(p)
MO1<- CX(DT,t)
FP1<- approxfun(c(0,1-MO1$TN,1),c(0,MO1$TP,1))(p)
MO2<- KM(DT,t)
FP2<- approxfun(c(0,1-MO2$TN,1),c(0,MO2$TP,1))(p)
Ekm[b,1]<- n^0.5*0.01*sum(abs(FR1-FK))
Eak[b,1]<- n^0.5*0.01*sum(abs(FR1-FA))
Edi[b,1]<- n^0.5*0.01*sum(abs(FR1-FD))
Eco[b,1]<- n^0.5*0.01*sum(abs(FR1-CO))
Ep1[b,1]<- n^0.5*0.01*sum(abs(FR1-FP1))
Ep2[b,1]<- n^0.5*0.01*sum(abs(FR1-FP2))

print(c(b,1,mean(Ekm[1:b,1]),mean(Eak[1:b,1]),mean(Edi[1:b,1]),mean(Ep1[1:b,1]),
       mean(Ep2[1:b,1])))

t<- 0
km<- survivalROC(Stime=exp(DT[,1]),status=DT[,2],marker = DT[,3],
                  predict.time = exp(t),method = "KM" )
FK<- approxfun(c(0,km$FP,1),c(0,km$TP,1))(p)
AK<- survivalROC(Stime=exp(DT[,1]),status=DT[,2],marker = DT[,3],
                  predict.time = exp(t),span = 0.1*n^(-0.20))
FA<- approxfun(c(0,AK$FP,1),c(0,AK$TP,1))(p)
DI<- rocDI(DT,t)
FD<- approxfun(c(0,DI$FP,1),c(0,DI$TP,1))(p)
```

```
MO1<- CX(DT,t)
FP1<- approxfun(c(0,1-MO1$TN,1),c(0,MO1$TP,1))(p)
MO2<- KM(DT,t)
FP2<- approxfun(c(0,1-MO2$TN,1),c(0,MO2$TP,1))(p)
Ekm[b,2]<- n^0.5*0.01*sum(abs(FR2-FK))
Eak[b,2]<- n^0.5*0.01*sum(abs(FR2-FA))
Edi[b,2]<- n^0.5*0.01*sum(abs(FR2-FD))
Ep1[b,2]<- n^0.5*0.01*sum(abs(FR2-FP1))
Ep2[b,2]<- n^0.5*0.01*sum(abs(FR2-FP2))

print(c(b,2,mean(Ekm[1:b,2]),mean(Eak[1:b,2]),mean(Edi[1:b,2]),mean(Ep1[1:b,2]),
       mean(Ep2[1:b,2])))

t<- 1
km<- survivalROC(Stime=exp(DT[,1]),status=DT[,2],marker = DT[,3],
                  predict.time = exp(t),method = "KM" )
FK<- approxfun(c(0,km$FP,1),c(0,km$TP,1))(p)
AK<- survivalROC(Stime=exp(DT[,1]),status=DT[,2],marker = DT[,3],
                  predict.time = exp(t),span = 0.1*n^(-0.20))
FA<- approxfun(c(0,AK$FP,1),c(0,AK$TP,1))(p)
DI<- rocDI(DT,t)
FD<- approxfun(c(0,DI$FP,1),c(0,DI$TP,1))(p)
MO1<- CX(DT,t)
FP1<- approxfun(c(0,1-MO1$TN,1),c(0,MO1$TP,1))(p)
MO2<- KM(DT,t)
FP2<- approxfun(c(0,1-MO2$TN,1),c(0,MO2$TP,1))(p)
Ekm[b,3]<- n^0.5*0.01*sum(abs(FR3-FK))
Eak[b,3]<- n^0.5*0.01*sum(abs(FR3-FA))
Edi[b,3]<- n^0.5*0.01*sum(abs(FR3-FD))
Ep1[b,3]<- n^0.5*0.01*sum(abs(FR3-FP1))
Ep2[b,3]<- n^0.5*0.01*sum(abs(FR3-FP2))

print(c(b,3,mean(Ekm[1:b,3]),mean(Eak[1:b,3]),mean(Edi[1:b,3]),mean(Ep1[1:b,3]),
       mean(Ep2[1:b,3])))
}
```

Capítulo 5

Aplicación a datos reales

En este capítulo, se aplica la metodología propuesta a una base de datos real. El conjunto de datos pertenece al estudio realizado por COCOMICS (*COllaborative COhorts to assess Multicomponent Indices of COPD in Spain*). Para el lector interesado en la obtención de información sobre la base de datos se proponen dos referencias: *Soriano y otros* [30] y *Marin y otros* [18].

Lo que se pretende estudiar es la capacidad del marcador FEV_1 (volumen espirado máximo en el primer segundo de la espiración forzada, que supone una medida de flujo del aire espirado) para predecir la mortalidad de pacientes de EPOC (enfermedad pulmonar obstructiva crónica). Este estudio es importante, ya que la EPOC está considerada como una enfermedad frecuente, prevenible y tratable, caracterizada por la limitación de la capacidad de introducir aire en los pulmones. Generalmente es progresiva y está asociada con una respuesta de las vías respiratorias y los pulmones ante gases y partículos nocivas presentes en el aire. La espirometría, que consiste en una serie de pruebas respiratorias que miden las capacidades y volúmenes pulmonares, así como la rapidez con que éstos pueden ser movilizados, está reconocida en guías clínicas como la prueba fundamental para el diagnóstico de EPOC, así como su gravedad y la etapa en la que se encuentra ésta.

Esta base de datos incluye 11 cohortes (grupos de pacientes estudiados en el mismo intervalo de tiempo) de EPOC recogidas en varias zonas de España, con un total de 3633

pacientes, de los cuales se han extraído los valores de una serie de biomarcadores, útiles en el tratamiento de la enfermedad. Entre todos ellos, el aquí estudiado es el ya mencionado FEV_1.

Estamos por tanto ante un problema que podemos tratar aplicando las diferentes propuestas citadas en este trabajo. Con el fin de hacernos una idea de cómo es la función de supervivencia, independientemente del valor del biomarcador, se ha realizado una estimación de $S(t)$ mediante el tradicional método de Kaplan-Meier, obteniendo los siguientes resultados:

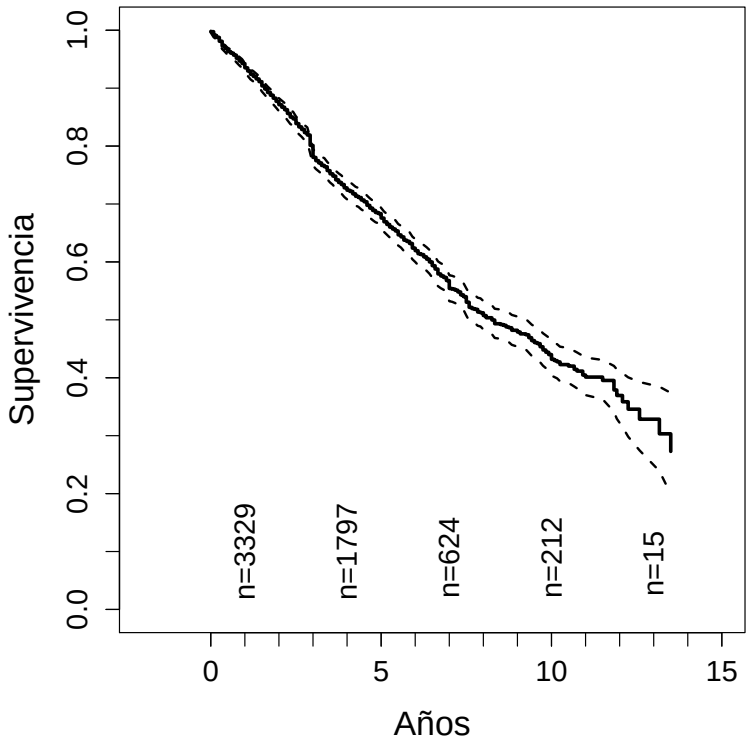


Figura 5.1: Estimación de la función de supervivencia para la base de datos de COCOMICS

En esta figura se ha añadido el número de pacientes en riesgo en los siguientes instantes de tiempo t: 1, 4, 7, 10 y 13 años. Además, se muestra el intervalo de confianza de la función de supervivencia al 95 %.

Veamos ahora las estimaciones de la curva ROC acumulativa/dinámica a partir de las

propuestas que vienen utilizándose hasta el momento, siguiendo con la misma notación. Se han estimado sobre los distintos tiempos t considerados anteriormente, excepto $t = 1$.

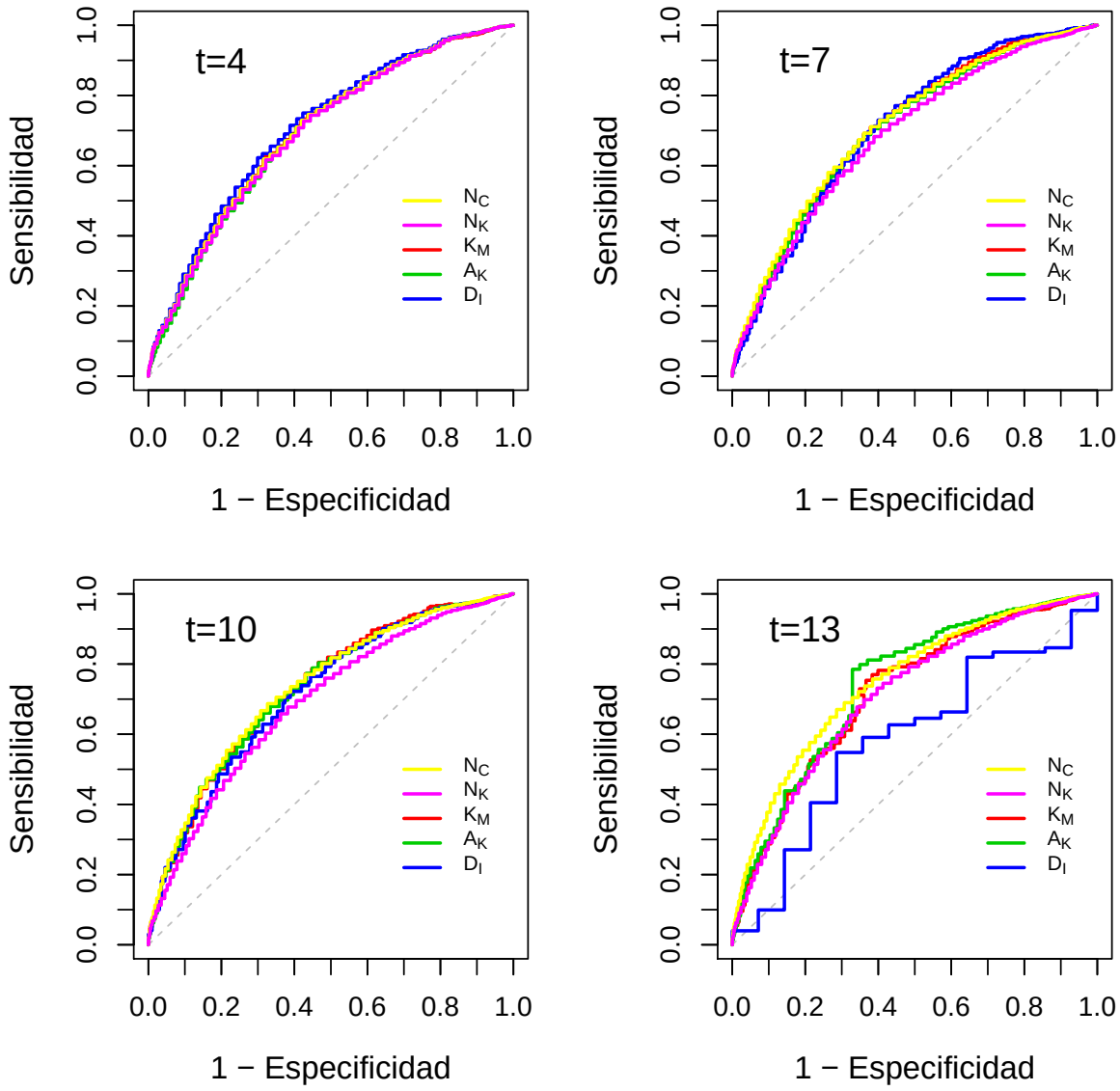


Figura 5.2: Estimación de $\mathcal{R}_t$ para diversos tiempos t (4, 7, 10, 13) según las diversas metodologías

Se observa que conforme el tiempo t considerado aumenta, las diferencias entre las distintas propuestas van aumentando. En cualquier caso, parece que el estimador nuevo propuesto basado en el método de Kaplan-Meier (N_K) es más conservativo que el resto.

Al estimar $\mathcal{R}_{t=13}$ se observa que la utilización por el método directo utilizando únicamente los sujetos no censurados (las otras cuatro propuestas sí tienen en cuenta estos datos) es el que peor aproximación aporta, lo cual se debe a que el número de individuos utilizados con diagnóstico negativo ($T > 13$) es pequeño.

Cabe mencionar también que en la figura de abajo a la derecha, la gráfica de la curva ROC estimada por el método del vecino más próximo (KNN) da un salto, lo cual es debido a la elección del parámetro de suavizado, que ha sido considerado $0.1 \cdot N^{-1/5}$ durante todo el trabajo. Esto pone de manifiesto nuevamente el gran inconveniente de utilizar este método y, además, que la elección de un parámetro de suavizado adecuado ya no sólo depende de la muestra considerada, sino también del tiempo t para el que estamos estimando la curva.

Por último, se ha calculado el área bajo la curva ROC a lo largo del tiempo t para las estimaciones de $\mathcal{R}_t$ utilizando el nuevo método basado en el modelo de Cox, con el objetivo cuantificar la calidad del biomarcador para predecir la mortalidad de estos pacientes. El resultado (AUC), junto con un intervalo de confianza al 95 % para dicha área, calculado a través de 100 réplicas bootstrap, se muestra en la siguiente figura.

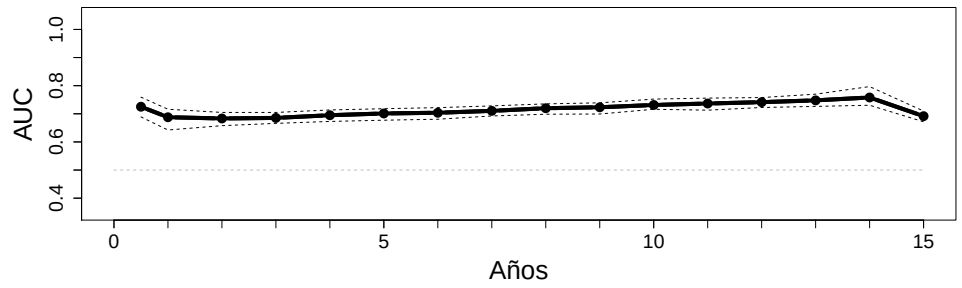


Figura 5.3: Evolución del área bajo la curva ROC utilizando N_K junto con un intervalo de confianza al 95 %

Como se puede observar, el mínimo lo alcanza al estimar $t = 2$ años, donde el AUC toma el valor 0.68. A partir de ese punto, el área bajo la curva va creciendo lentamente, alcanzando su máximo valor para $t = 14$ años, siendo este valor igual a 0.75.

Conclusiones

En este trabajo se ha realizado una revisión metodológica de diferentes propuestas para estimar la curva ROC tiempo-dependiente, en particular el enfoque acumulativo/dinámico de ésta, así como la incorporación de un nuevo método con el objeto de poner fin a los problemas que presentan el resto de propuestas. Con el fin de que todo aquel lector no especializado en el tema logre comprender la finalidad del mismo y pueda aplicarlo en casos prácticos si así lo requiere, en las primeras páginas se han recopilado una serie de conceptos y resultados básicos propios de las curvas ROC y del análisis de supervivencia.

La metodología propuesta asigna una probabilidad de pertenecer al grupo negativo/control (respectivamente al grupo positivo/caso) a aquellos pacientes cuyo estado real en el tiempo considerado es desconocido, es decir, aquellos que han sido censurados antes de la ocurrencia del evento de interés. Esta probabilidad puede estimarse, según esta propuesta, mediante dos métodos diferentes, uno semi-paramétrico (basado en la regresión de Cox considerando el valor del biomarcador como única covariable) y otro no paramétrico (basado en la estimación de Kaplan-Meier sobre un determinado subconjunto).

Sin embargo, se pueden utilizar otros métodos con el mismo objetivo, abarcando desde métodos adaptados a situaciones particulares como riesgos competitivos o en contextos multi-estado, hasta la generalización de la curva ROC en caso de no ser monótona [22]. El estimador propuesto soluciona los inconvenientes que presentaban los anteriores:

1. es monótono,

2. siempre toma valores entre 0 y 1, y

3. no depende de parámetros de suavizado.

Además, los estudios de simulación sugieren que tanto la nueva propuesta (utilizando ambas metodologías) funciona siempre mejor que las anteriores cuando el porcentaje de censura en la muestra es elevado. Cuando se considera que existe alguna correlación entre el biomarcador y el tiempo de censura, los resultados no cambian sustancialmente.

Cabe mencionar que el método directo, es decir, aquel obtenido eliminando los sujetos problemáticos de los que venimos hablando, normalmente proporciona buenos resultados. Sin embargo, el no considerar esos individuos no es una buena elección, ya que estaríamos eliminando un subconjunto de individuos que pueden tener propiedades interesantes. Por ejemplo, si existe algún tipo de relación entre el biomarcador considerado y el tiempo de censura, los resultados podrían ser muy diferentes de los reales.

Quizás, el principal defecto de la metodología presentada es la falta de un estudio riguroso de sus propiedades teóricas. Sin embargo, el estimador propuesto tiene dos partes diferentes: la primera es la curva ROC tradicional para datos completos, mientras que la segunda depende en gran medida de la capacidad para predecir la probabilidad de que un sujeto censurado antes del tiempo t considerado, sea un *control* (respectivamente un *caso*). Esta probabilidad depende de varios factores que no están claros *a priori*. En este sentido, conviene recalcar de nuevo los buenos resultados del estudio de simulación.

La continuación natural de este trabajo sería el estudio de las propiedades de los estimadores propuestos y la construcción de un paquete en el software estadístico `R` que facilite la utilización del estimador de $\mathcal{R}_t^{\mathbb{A}/\mathbb{D}}$ propuesto. Por otra parte, sería también muy interesante estudiar la aplicación del estimador de la probabilidad de pertenencia de estos sujetos "problemáticos" a cada grupo con otras finalidades distintas a la propuesta en este trabajo.

Anexos

Figura 1.1

```
x <- rnorm(500000,-1,3)
y <- rnorm(500000,3,4)

library(ROCR)

datos <- data.frame(marker=c(x,y),group=as.factor(c(rep(0,500000),rep(1,500000))))
attach(datos)

par(mfrow=c(1,2))

plot(density(x),col=2,xlim=c(-12,18),xlab="Puntos de corte",ylab="Funcion de densidad",
       main="Valores del biomarcador en ambos grupos")
lines(density(y),col=3)
legend(8.4,0.13,legend=c("Controles (D=0)","Casos (D=1)"),lty=c(1,1),col=c(2,3))
abline(v=-5,lty=2,col=4,lwd=2)
abline(v=1.525,lty=2,col=6,lwd=2)
abline(v=7,lty=2,col=7,lwd=2)

pred <- prediction(datos$marker, datos$group)
perf <- performance(pred,"tpr","fpr")
plot(perf,print.cutoffs.at=c(-5.0,1.52,7.0),text.adj=c(1,-1), avg="threshold", lwd=3,
      xlab="1-Especificidad",ylab="Sensibilidad",main="Curva ROC")
lines(c(0,1),c(0,1))
points(c(0.005,0.2,0.91),c(0.16,0.645,0.978),lwd=4,col=c(7,6,4))
```

Figura 1.2

```
x <- rnorm(500000,-5,1.5)
y <- rnorm(500000,3,1.2)

datos <- data.frame(marker=c(x,y),group=as.factor(c(rep(0,500000),rep(1,500000))))
attach(datos)

par(mfrow=c(1,2))

plot(density(x),col=2,xlim=c(-13,18),ylim=c(0,0.35),xlab="Valores del biomarcador X",
      ylab="Funcion de densidad",main="Distribucion del biomarcador en ambos grupos")
lines(density(y),col=3)
legend(7.4,0.3,legend=c("Controles (D=0)","Casos (D=1)"),lty=c(1,1),col=c(2,3))

pred <- prediction(datos$marker, datos$group)
perf <- performance(pred,"tpr","fpr")
plot(perf,lwd=2,xlab="1-Especificidad",ylab="Sensibilidad",main="Curva ROC")
lines(c(0,1),c(0,1))
performance(pred,"auc")
legend(0.63,0.16,legend="AUC = 0.99998")

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

x <- rnorm(500000,-4,3)
y <- rnorm(500000,3,4)

datos <- data.frame(marker=c(x,y),group=as.factor(c(rep(0,500000),rep(1,500000))))
attach(datos)

par(mfrow=c(1,2))

plot(density(x),col=2,xlim=c(-17,20),xlab="Valores del biomarcador X",ylab="Funcion
      de densidad",main="Distribucion del biomarcador en ambos grupos")
lines(density(y),col=3)
```

```
legend(6,0.13,legend=c("Controles (D=0)","Casos (D=1)"),lty=c(1,1),col=c(2,3))

pred <- prediction(datos$marker, datos$group)
perf <- performance(pred,"tpr","fpr")
plot(perf,lwd=2,xlab="1-Especificidad",ylab="Sensibilidad",main="Curva ROC")
lines(c(0,1),c(0,1))
performance(pred,"auc")
legend(0.63,0.16,legend="AUC = 0.91942")

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

x <- rnorm(500000,-2,3.1)
y <- rnorm(500000,1,4)

datos <- data.frame(marker=c(x,y),group=as.factor(c(rep(0,500000),rep(1,500000))))
attach(datos)

par(mfrow=c(1,2))

plot(density(x),col=2,xlim=c(-17,20),ylim=c(0,0.14),xlab="Valores del biomarcador X",
      ylab="Funcion de densidad",main="Distribucion del biomarcador en ambos grupos")
lines(density(y),col=3)
legend(6,0.13,legend=c("Controles (D=0)","Casos (D=1)"),lty=c(1,1),col=c(2,3))

pred <- prediction(datos$marker, datos$group)
perf <- performance(pred,"tpr","fpr")
plot(perf,lwd=2,xlab="1-Especificidad",ylab="Sensibilidad",main="Curva ROC")
lines(c(0,1),c(0,1))
performance(pred,"auc")
legend(0.63,0.16,legend="AUC = 0.72318")
```

Figura 1.3

```
tiempos <- c(13,17,20,20,24,32,36,45)
```

```
censuras <- rep(1,8)
datos <- data.frame(tiempos,censuras)

library(survival)

S <- survfit(Surv(tiempos,censuras)~1,conf.type="log",conf.int=0,type="kaplan-meier",
             error="greenwood",data=datos)
plot(S,lwd=2,mark.time=TRUE,xlab="Meses",ylab="S(t)",main="Curva de supervivencia")
```

Figura 1.4

```
plot(-1,-1,xlim=c(0,1),ylim=c(0,0.5),frame=F,axes=F,ylab=" ",xlab=" ",cex.lab=2)

axis(1,at = 0.65, lab ="Fin del estudio",cex.lab=2.5,tick=F)
axis(1,at = 0.1, lab ="Inicio del estudio",cex.lab=2.5,tick=F)

lines(c(0.65,0.65),c(0,1),lty=2,lwd=2)
lines(c(0.1,0.1),c(0,1),lty=2,lwd=2)

lines(c(0.1,0.6),c(0.1,0.1),lwd=5,type="l")
points(0.6,0.1,pch=1,lwd=5)
lines(c(0.1,0.5),c(0.2,0.2),lwd=5,type="l")
points(0.5,0.2,pch=4,lwd=5)
lines(c(0.2,0.75),c(0.3,0.3),lwd=5,type="l")
points(0.65,0.3,pch=1,lwd=4)
points(0.76,0.3,pch=4,lwd=5)
lines(c(0.15,0.9),c(0.4,0.4),lwd=5,type="l")
points(0.65,0.4,pch=1,lwd=4)
points(0.9,0.4,pch=4,lwd=5)

text(0,0.1,"D",cex=1.5)
text(0,0.2,"C",cex=1.5)
text(0,0.3,"B",cex=1.5)
text(0,0.4,"A",cex=1.5)
```

Figura 1.5

```
tiempos <- c(13,14,15,16,20,24,28,34)
censuras <- c(1,0,0,1,1,0,1,1)
datos <- data.frame(tiempos,censuras)

library(survival)

S <- survfit(Surv(tiempos,censuras)~1,conf.type="log",conf.int=0,type="kaplan-meier",
             error="greenwood",data=datos)
plot(S,lwd=2,mark.time=TRUE,xlab="Meses",ylab="S(t)",main="Curva de supervivencia")
```

Figura 1.6

```
library(survival)
par(mfrow=c(1,2))

tiempos <- c(13,17,20,20,24,32,36,45)
censuras <- rep(1,8)
datos <- data.frame(tiempos,censuras)

S <- survfit(Surv(tiempos,censuras)~1,conf.type="plain",conf.int=0.95,type="kaplan-meier",
             error="greenwood",data=datos)
plot(S,lwd=2,mark.time=TRUE,xlab="Meses",ylab="S(t)",main="Curva de supervivencia")

tiempos <- c(13,14,15,16,20,24,28,34)
censuras <- c(1,0,0,1,1,0,1,1)
datos <- data.frame(tiempos,censuras)

S <- survfit(Surv(tiempos,censuras)~1,conf.type="plain",conf.int=0.95,type="kaplan-meier",
             error="greenwood",data=datos)
plot(S,lwd=2,mark.time=TRUE,xlab="Meses",ylab="S(t)",main="Curva de supervivencia")
```

Figura 1.7

```
library(survival)

hmohiv<-read.table("http://www.ats.ucla.edu/stat/r/examples/asa/hmohiv.csv", sep=",",
                   header=TRUE)
attach(hmohiv)

cox1 <- coxph( Surv(time,censor)~age+drug,na.action=na.exclude)
summary(cox1)
plot(survfit(cox1),conf.int=FALSE,main="",xlab="Meses",ylab="S(t)",lty=1,lwd=2)

S <- survfit(Surv(time,censor)~1,conf.type="plain",conf.int=FALSE,type="kaplan-meier",
             error="greenwood",na.action=na.exclude,data=hmohiv)
lines(S,lwd=2,lty=2,mark.time=TRUE,xlab="Meses",ylab="S(t)",main="Curva de supervivencia")

legend(25,0.9,legend=c("Ajuste por Cox","Estimador Kaplan-Meier"),lty=c(1,2),lwd=2)
```

Figura 2.1

```
plot(-1,-1,xlim=c(0,1),ylim=c(0,0.5),frame=F,axes=F,ylab=" ",xlab=" ",cex.lab=2)

axis(1,at = 0.75, lab ="t*",cex.lab=2.5,tick=F)
axis(1,at = 0.35, lab ="t",cex.lab=2.5,tick=F)
axis(1,at = 0.1, lab ="Inicio seguimiento \n de cada sujeto",cex.lab=0.6,tick=F)
lines(c(0.875,0.875),c(0,1),lwd=125,col="light green")
lines(c(0.35,0.35),c(0,1),lwd=10,col="light blue")

lines(c(0.1,0.1),c(0,1),lty=2,lwd=2)
lines(c(0.75,0.75),c(0,1),lty=2,lwd=2)
lines(c(0.35,0.35),c(0,1),lty=2,lwd=2)

lines(c(0.1,0.5),c(0.1,0.1),lwd=5,type="l")
points(0.5,0.1,pch=4,lwd=5)
text(0.02,0.1,"No se \n considera",cex=0.7)
```

```
lines(c(0.1,0.35),c(0.2,0.2),lwd=5,type="l")
points(0.35,0.2,pch=4,lwd=5)
text(0.02,0.2,"Caso",cex=0.7)

lines(c(0.1,0.85),c(0.3,0.3),lwd=5,type="l")
points(0.85,0.3,pch=4,lwd=5)
text(0.02,0.3,"Control",cex=0.7)

lines(c(0.1,0.2),c(0.4,0.4),lwd=5,type="l")
points(0.2,0.4,pch=4,lwd=5)
text(0.02,0.4,"No se \n considera",cex=0.7)
```

Figura 2.2

```
plot(-1,-1,xlim=c(0,1),ylim=c(0,0.5),frame=F,axes=F,ylab=" ",xlab=" ",cex.lab=2)

axis(1,at = 0.6, lab ="t",cex.lab=2.5,tick=F)
axis(1,at = 0.1, lab ="Inicio seguimiento \n de cada sujeto",cex.lab=0.6,tick=F)
lines(c(0.6,0.6),c(0,1),lwd=10,col="light blue")
lines(c(0.8,0.8),c(0,1),lwd=200,col="light green")

lines(c(0.1,0.1),c(0,1),lty=2,lwd=2)
lines(c(0.6,0.6),c(0,1),lty=2,lwd=2)

lines(c(0.1,0.5),c(0.1,0.1),lwd=5,type="l")
points(0.5,0.1,pch=4,lwd=5)
text(0.02,0.1,"No se \n considera",cex=0.7)

lines(c(0.1,0.6),c(0.2,0.2),lwd=5,type="l")
points(0.6,0.2,pch=4,lwd=5)
text(0.02,0.2,"Caso",cex=0.7)

lines(c(0.1,0.85),c(0.3,0.3),lwd=5,type="l")
```

```
points(0.85,0.3,pch=4,lwd=5)
text(0.02,0.3,"Control",cex=0.7)

lines(c(0.1,0.2),c(0.4,0.4),lwd=5,type="l")
points(0.2,0.4,pch=4,lwd=5)
text(0.02,0.4,"No se \n considera",cex=0.7)
```

Figura 2.3

```
plot(-1,-1,xlim=c(0,1),ylim=c(0,0.5),frame=F,axes=F,ylab=" ",xlab=" ",cex.lab=2)

axis(1,at = 0.6, lab ="t",cex.lab=2.5,tick=F)
axis(1,at = 0.1, lab ="Inicio seguimiento \n de cada sujeto",cex.lab=0.6,tick=F)
lines(c(0.8,0.8),c(0,1),lwd=200,col="light green")
lines(c(0.35,0.35),c(0,1),lwd=250,col="light blue")

lines(c(0.1,0.1),c(0,1),lty=2,lwd=2)
lines(c(0.6,0.6),c(0,1),lty=2,lwd=2)

lines(c(0.1,0.5),c(0.1,0.1),lwd=5,type="l")
points(0.5,0.1,pch=4,lwd=5)
text(0.02,0.1,"Caso",cex=0.7)

lines(c(0.1,0.6),c(0.2,0.2),lwd=5,type="l")
points(0.6,0.2,pch=4,lwd=5)
text(0.02,0.2,"Caso",cex=0.7)

lines(c(0.1,0.85),c(0.3,0.3),lwd=5,type="l")
points(0.85,0.3,pch=4,lwd=5)
text(0.02,0.3,"Control",cex=0.7)

lines(c(0.1,0.2),c(0.4,0.4),lwd=5,type="l")
points(0.2,0.4,pch=4,lwd=5)
text(0.02,0.4,"Caso",cex=0.7)
```

Figura 3.1

```
plot(-1,-1,xlim=c(0,1),ylim=c(0,0.5),frame=F,axes=F,ylab=" ",xlab=" ",cex.lab=1)
lines(c(0.85,0.85),c(0,1),lwd=200,col="light gray")

axis(1,at = 0.1, lab ="Inicio seguimiento \n de cada sujeto",cex.lab=0.5,tick=F)
axis(1,at = 0.65, lab ="t",cex.lab=2.5,tick=F)
lines(c(0.1,0.1),c(0,1),lty=2,lwd=2)
lines(c(0.65,0.65),c(0,1),lty=2,lwd=2)

lines(c(0.1,0.5),c(0.2,0.2),lwd=5,type="l")
points(0.5,0.2,pch=4,lwd=5)

lines(c(0.1,0.75),c(0.3,0.3),lwd=5,type="l")
points(0.76,0.3,pch=1,lwd=4)

lines(c(0.1,0.9),c(0.4,0.4),lwd=5,type="l")
points(0.9,0.4,pch=4,lwd=5)

lines(c(0.1,0.6),c(0.1,0.1),lwd=5,type="l")
points(0.61,0.1,pch=1,lwd=5)

text(0,0.1,"D",cex=1)
text(0,0.2,"C",cex=1)
text(0,0.3,"B",cex=1)
text(0,0.4,"A",cex=1)
```

Figura 3.2

```
plot(-1,-1,xlim=c(0,1),ylim=c(0,0.5),frame=F,axes=F,ylab=" ",xlab=" ",cex.lab=2)

axis(1,at = 0.1, lab ="Inicio seguimiento \n de cada sujeto",cex.lab=0.6,tick=F)
axis(1,at = 0.3, lab ="t=1",cex.lab=2.5,tick=F)
axis(1,at = 0.5, lab ="t=2",cex.lab=2.5,tick=F)
axis(1,at = 0.7, lab ="t=3",cex.lab=2.5,tick=F)
```

```
axis(1,at = 0.9, lab ="t=4",cex.lab=2.5,tick=F)
axis(2.7,at = 0.25, lab ="Valores del biomarcador",cex.lab=0.5,tick=F)

lines(c(0.1,0.1),c(0,1),lty=2,lwd=2)
lines(c(0.3,0.3),c(0,1),lty=2,lwd=2)
lines(c(0.5,0.5),c(0,1),lty=2,lwd=2)
lines(c(0.7,0.7),c(0,1),lty=2,lwd=2)
lines(c(0.9,0.9),c(0,1),lty=2,lwd=2)

lines(c(0.1,0.4),c(0.1,0.1),lwd=5,type="l")
points(0.4,0.1,pch=1,lwd=8)
text(0.02,0.1,"x=4",cex=1.2)

lines(c(0.1,0.9),c(0.2,0.2),lwd=5,type="l")
points(0.9,0.2,pch=4,lwd=5)
text(0.02,0.2,"x=3.6",cex=1.2)

lines(c(0.1,0.3),c(0.3,0.3),lwd=5,type="l")
points(0.3,0.3,pch=4,lwd=5)
text(0.02,0.4,"x=1",cex=1.2)

lines(c(0.1,0.5),c(0.4,0.4),lwd=5,type="l")
points(0.5,0.4,pch=4,lwd=5)
text(0.02,0.3,"x=2.1",cex=1.2)
```

Figura 3.3

```
library(survivalROC)
library(risksetROC)
library(KMsurv)
library(mvtnorm)
library(ks)
library(pROC)
```

```
data(kidtran)

####  COX

CX<- function(DT,t)
{
prob<- NULL
T<- DT[,1]; E<- DT[,2]; M<- DT[,3] # Para cada individuo: T=tiempo observado, E=delta
                                   # (1 si el dato es completo, 0 si es censurado),
                                   # M=valor del biomarcador X
P<- which(T <= t & E==1) # Indices de los casos
N<- which(T > t) # Indices de los controles
IN<- which(T <= t & E==0) # Indices de los datos censurados antes de t

if (length(IN)>0) # Si hay datos censurados antes de t (undefined subjects)
{fit <- coxph(Surv(T, E) ~ M)} # Ajuste del modelo de Cox con covariable M (biomarcador)

prob<- 1:length(IN) # Vector de probabilidades de longitud el nUmero de datos censurados
                    # antes de t

for (j in 1:length(IN))
{
#pos <- which(M==M[IN[j]])
md<- survfit(fit, newdata=data.frame(M=M[IN[j]]))
f<- approxfun(c(0,md$time),c(1,md$surv))
prob[j]<- f(t)/f(T[IN[j]])
if (is.na(prob[j])) prob[j]<- 1
}
cut<- c(min(unique(sort(M)))-1,unique(sort(M))) #  Vector que contiene los puntos de
                                                # corte c a considerar
nS<- length(P) + sum(1-prob) # Numero de casos + Suma probabilidades de que los
                                                 # censurados antes de t sean casos
                             # Denominador sensibilidad
nE<- length(N) + sum(prob) # Numero de controles + Suma probabilidades de que los
                                                 # censurados antes de t sean controles
```

```
                              # Denominador especificidad
Se<- cut; Es<- cut
for (i in 1:length(cut))
{
Se[i]<- (sum(M[P]> cut[i]) + sum(1-prob[which(M[IN]>cut[i])]))/nS
Es[i]<- (sum(M[N]<= cut[i]) + sum(prob[which(M[IN]<=cut[i])]))/nE
}
list(TP=Se, TN=Es, prob=prob)
}

####  KAPLAN-MEIER

KM<- function(DT,t)
{
prob<- NULL
T<- DT[,1]; E<- DT[,2]; M<- DT[,3]
P<- which(T <= t & E==1)
N<- which(T > t)
IN<- which(T <= t & E==0)

if (length(IN)>0)
{
prob<- 1:length(IN)
for (j in 1:length(IN))
{
I<- which(M<= M[IN[j]])
fit<- survfit(Surv(T[I],E[I])~1)
f<- stepfun(fit$time,c(1,fit$surv))
prob[j]<- f(t)/f(T[IN[j]])
if (is.na(prob[j])) prob[j]<- 1
}
}
cut<- c(-5,as.numeric(names(table(M))))
nS<- length(P) + sum(1-prob)
nE<- length(N) + sum(prob)
```

```
Se<- cut; Es<- cut
for (i in 1:length(cut))
{
Se[i]<- (sum(M[P]> cut[i]) + sum(1-prob[which(M[IN]>cut[i])]))/nS
Es[i]<- (sum(M[N]<= cut[i]) + sum(prob[which(M[IN]<=cut[i])]))/nE
}
print(c(min(f(T[IN]),max(f(T[IN]))

list(TP=Se, TN=Es, prob=prob)
}

###### ESTIMACION A LOS 9 ANOS

tt<- 9*365.25

KMfallo<- survivalROC(Stime=kidtran$time,status=kidtran$delta,marker = kidtran$age,
                      predict.time = tt,method = "KM" )
nobs<- length(kidtran$time)
K1<- survivalROC(Stime=kidtran$time,status=kidtran$delta,marker = kidtran$age,
                      predict.time = 1825,method = "NNE",,span = 0.01*nobs^(-0.20) )

P<- which(kidtran$time< tt & kidtran$delta==1)
N<- which(kidtran$time>= tt)

di<- roc(c(rep(1,length(P)),rep(0,length(N))),c(kidtran$age[P],kidtran$age[N]))

########

cx<- CX(cbind(kidtran$time,kidtran$delta,kidtran$age),tt)
km<- KM(cbind(kidtran$time,kidtran$delta,kidtran$age),tt)

par(mfrow=c(1,2))

plot(KMfallo$FP, KMfallo$TP,type="l",lwd=3,xlab="1 - Especificidad",
ylab="Sensibilidad",cex.lab=1.25,xlim=c(0,1),ylim=c(0,1),cex.axis=1,col="gray")
```

```
lines(c(0,1),c(1,1),col="gray",lty=2)
lines(c(0,1),c(0,1),col="gray",lty=2)
lines(K1$FP,K1$TP,lwd=3,col="gray",lty=2)
lines(1-di$spec,di$sens,lwd=3)
lines(c(0.6,0.7),c(0.4,0.4),lwd=3)
text(0.72,0.4,expression(D[I]),cex=1.25,pos=4)
lines(c(0.6,0.7),c(0.33,0.33),lwd=3,col="gray")
text(0.72,0.33,expression(K[M]),cex=1.25,pos=4)
lines(c(0.6,0.71),c(0.26,0.26),lwd=3,col="gray",lty=2)
text(0.72,0.26,expression(A[K]),cex=1.25,pos=4)
axis(1,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
axis(2,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))

plot(1-cx$TN, cx$TP,type="l",lwd=3,xlab="1 - Especificidad",
ylab="Sensibilidad",cex.lab=1.25,xlim=c(0,1),ylim=c(0,1),cex.axis=1,col="gray")
lines(c(0,1),c(1,1),col="gray",lty=2)
lines(c(0,1),c(0,1),col="gray",lty=2)
lines(1- km$TN,km$TP,lwd=3,col="gray",lty=2)
lines(1-di$spec,di$sens,lwd=3)
lines(c(0.6,0.7),c(0.4,0.4),lwd=3)
text(0.72,0.4,expression(D[I]),cex=1.25,pos=4)
lines(c(0.6,0.7),c(0.33,0.33),lwd=3,col="gray")
text(0.72,0.33,expression(N[C]),cex=1.25,pos=4)
lines(c(0.6,0.71),c(0.26,0.26),lwd=3,col="gray",lty=2)
text(0.72,0.26,expression(N[K]),cex=1.25,pos=4)
axis(1,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
axis(2,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
```

Figura 4.1

```
library(survivalROC)
library(risksetROC)
library(KMsurv)
library(mvtnorm)
```

```
library(ks)
library(pROC)

c<- -0.25 # Coeficiente de correlacion rho = -1/4
z<- seq(-5,5,0.1) # Puntos de corte c a considerar para construir la grafica
p<- seq(0,1,0.01) # Secuencia de puntos para los que se representara R(p)

#####

mu<- c(0,0) # Media del vector aleatorio (log(T),X)
sigma<- matrix(c(1,c,c,1),ncol=2,nrow=2) # Matriz de varianzas-covarianzas del v.a.
                                         # (log(T),X)

# Nota: la funcion pmvnorm(inf,sup,media,sigma) calcula la probabilidad
       # P(inf < Y < sup) siendo Y un vector aleatorio normal con media = "media"
       # y matriz de varianzas-covarianzas = "sigma"
        # Almacena dicho valor en su primera componente [1]

## Curva ROC para log(t) = -1

t<- -1
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),Inf,mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-Inf),c(Inf,t),mu,sigma)[1]/pnorm(t)
FR1<- approxfun(c(0,FP,1),c(0,TP,1))(p)

## Curva ROC para log(t) = 0

t<- 0
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),Inf,mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-Inf),c(Inf,t),mu,sigma)[1]/pnorm(t)
FR2<- approxfun(c(0,FP,1),c(0,TP,1))(p)
```

```
## Curva ROC para log(t) = 1

t<- 1
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),Inf,mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-Inf),c(50,t),mu,sigma)[1]/pnorm(t)
FR3<- approxfun(c(0,FP,1),c(0,TP,1))(p)

## Grafica

par(mfrow=c(1,2))
plot(p,FR1,type="l",lwd=2,xlab="1 - Especificidad",
ylab="Sensibilidad",cex.lab=1.6,xlim=c(0,1),ylim=c(0,1),cex.axis=1.5,col=4)
lines(c(0,1),c(0,1),col=1,lty=2)
lines(p,FR2,lwd=2,col=6)
lines(p,FR3,lwd=2,col=3)
lines(c(0.6,0.7),c(0.4,0.4),lwd=3,col=4)
text(0.72,0.4,"log(t) = -1",cex=1.5,pos=4)
lines(c(0.6,0.7),c(0.33,0.33),lwd=3,col=6)
text(0.72,0.33,"log(t) = 0",cex=1.5,pos=4)
lines(c(0.6,0.7),c(0.26,0.26),lwd=3,col=3)
text(0.72,0.26,"log(t) = 1",cex=1.5,pos=4)
text(0.03,0.95,expression(rho*" = -1/4"),cex=1.5,pos=4)
axis(1,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
axis(2,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))

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

c<- -0.75 # Coeficiente de correlacion rho = -3/4
z<- seq(-5,5,0.1) # Puntos de corte c a considerar para construir la grafica
p<- seq(0,1,0.01) # Secuencia de puntos para los que se representara R(p)

#####
```

```
mu<- c(0,0) # Media del vector aleatorio (log(T),X)
sigma<- matrix(c(1,c,c,1),ncol=2,nrow=2) # Matriz de varianzas-covarianzas del v.a.
                                         # (log(T),X)

## Curva ROC para log(t) = -1

t<- -1
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),c(50,50),mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-50),c(50,t),mu,sigma)[1]/pnorm(t)
FR1<- approxfun(c(0,FP,1),c(0,TP,1))(p)

## Curva ROC para log(t) = 0

t<- 0
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),c(50,50),mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-50),c(50,t),mu,sigma)[1]/pnorm(t)
FR2<- approxfun(c(0,FP,1),c(0,TP,1))(p)

## Curva ROC para log(t) = 1

t<- 1
FP<- rep(-1,length(z))
for (i in 1:length(z)) FP[i]<- pmvnorm(c(z[i],t),c(50,50),mu,sigma)[1]/(1-pnorm(t))
TP<- rep(-1,length(z))
for (i in 1:length(z)) TP[i]<- pmvnorm(c(z[i],-50),c(50,t),mu,sigma)[1]/pnorm(t)
FR3<- approxfun(c(0,FP,1),c(0,TP,1))(p)

## Grafica

plot(p,FR1,type="l",lwd=2,xlab="1 - Especificidad",
```

```
ylab="Sensibilidad",cex.lab=1.6,xlim=c(0,1),ylim=c(0,1),cex.axis=1.5,col=4)
lines(c(0,1),c(0,1),col=1,lty=2)
lines(p,FR2,lwd=2,col=6)
lines(p,FR3,lwd=2,col=3)
lines(c(0.6,0.7),c(0.4,0.4),lwd=3,col=4)
text(0.72,0.4,"log(t) = -1",cex=1.5,pos=4)
lines(c(0.6,0.7),c(0.33,0.33),lwd=3,col=6)
text(0.72,0.33,"log(t) = 0",cex=1.5,pos=4)
lines(c(0.6,0.7),c(0.26,0.26),lwd=3,col=3)
text(0.72,0.26,"log(t) = 1",cex=1.5,pos=4)
text(0.03,0.95,expression(rho*" = -3/4"),cex=1.5,pos=4)
axis(1,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
axis(2,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
```

Figura 5.1

```
#### CONJUNTO DE DATOS

CCCs<- read.delim2("CCCs.txt")
attach(CCCs)
DT<- cbind(time,Exitus,FEV1)

###### SUPERVIVENCIA

md<- summary(survfit(Surv(time,Exitus==1)~1))

plot(c(md$time,max(time)),c(md$surv,min(md$surv)),type="s",lwd=3,xlab="Anos",
      ylab="Supervivencia",cex.lab=1.5,ylim=c(0,1),cex.axis=1.3,xlim=c(-2,15))
sum(sort(DT[,1])>=1)
text(1,0.1,"n=3329",cex=1.25,srt=90)
sum(sort(DT[,1])>=4)
text(4,0.1,"n=1797",cex=1.25,srt=90)
sum(sort(DT[,1])>=7)
text(7,0.09,"n=624",cex=1.25,srt=90)
```

```
sum(sort(DT[,1])>=10)
text(10,0.09,"n=212",cex=1.25,srt=90)
sum(sort(DT[,1])>=13)
text(13,0.08,"n=15",cex=1.25,srt=90)
axis(1,seq(0,15,1),lab=rep(" ",length(seq(0,15,1))))
axis(2,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))

lines(md$time,md$low,lty=2,lwd=2)
lines(md$time,md$up,lty=2,lwd=2)
```

Figura 5.2

```
library(survivalROC)
library(risksetROC)
library(KMsurv)
library(mvtnorm)
library(ks)
library(pROC)

## COX

CX<- function(DT,t)
{
prob<- NULL
T<- DT[,1]; E<- DT[,2]; M<- DT[,3]
P<- which(T <= t & E==1)
N<- which(T > t)
IN<- which(T <= t & E==0)

if (length(IN)>0)
{
fit <- coxph(Surv(T, E) ~ M)
md<- survfit(fit, newdata=data.frame(DT))
prob<- 1:length(IN)
```

```
for (j in 1:length(IN))
{
f<- approxfun(c(0,md$time),c(1,md$surv[,IN[j]]))
prob[j]<- f(t)/f(T[IN[j]])
if (is.na(prob[j])) prob[j]<- 1
}
}
cut<- sort(c(min(as.numeric(names(table(M))))-1,as.numeric(names(table(M)))))
nS<- length(P) + sum(1-prob)
nE<- length(N) + sum(prob)
Se<- cut; Es<- cut
for (i in 1:length(cut))
{
Se[i]<- (sum(M[P]> cut[i]) + sum(1-prob[which(M[IN]>cut[i])]))/nS
Es[i]<- (sum(M[N]<= cut[i]) + sum(prob[which(M[IN]<=cut[i])]))/nE
}
list(TP=Se, TN=Es, prob=prob)
}

##  KAPLAN-MEIER

KM<- function(DT,t)
{
prob<- NULL
T<- DT[,1]; E<- DT[,2]; M<- DT[,3]
P<- which(T <= t & E==1)
N<- which(T > t)
IN<- which(T <= t & E==0)

if (length(IN)>0)
{
prob<- 1:length(IN)
for (j in 1:length(IN))
{
I<- which(M<= M[IN[j]])
```

```
fit<- survfit(Surv(T[I],E[I])~1)
f<- stepfun(fit$time,c(1,fit$surv))
prob[j]<- f(t)/f(T[IN[j]])
if (is.na(prob[j])) prob[j]<- 1
}
}
cut<- unique(sort(M))
nS<- length(P) + sum(1-prob)
nE<- length(N) + sum(prob)
Se<- cut; Es<- cut
for (i in 1:length(cut))
{
Se[i]<- (sum(M[P]> cut[i]) + sum(1-prob[which(M[IN]>cut[i])]))/nS
Es[i]<- (sum(M[N]<= cut[i]) + sum(prob[which(M[IN]<=cut[i])]))/nE
}
list(TP=Se, TN=Es, prob=prob)
}

##### ROC sin datos censurados

rocDI <- function(DT,t)
{
I1<- which(DT[,1]<= t & DT[,2]==1)
I0<- which(DT[,1]>t)
if (length(I1)*length(I0)==0) {FP<- c(0,0.5,1); TP<- c(0,0.5,1)}
else {
M<- c(DT[I0,3],DT[I1,3])
R<- c(rep(0,length(I0)),rep(1,length(I1)))
r<- roc(R,M)
FP<- 1-r$spec; TP<- r$sens}
list( FP= FP, TP=TP)
}
```

```
##### CONJUNTO DE DATOS

CCCs<- read.delim2("CCCs.txt")
attach(CCCs)
DT<- cbind(time,Exitus,FEV1)

##### CURVAS ROC A LOS 4, 7, 10 Y 13 ANOS

I<- which(time>=0 & Exitus>=0 & FEV1>=0)
DTC<- cbind(DT[I,1],DT[I,2],-DT[I,3])

par(mfrow=c(2,2))

for(t in c(4,7,10,13)){
km<- survivalROC(Stime=DTC[,1],status=DTC[,2],marker = DTC[,3],predict.time = t ,
                  method = "KM" )
ak<- survivalROC(Stime=DTC[,1],status=DTC[,2],marker = DTC[,3],predict.time = t,
                  span = 0.1*length(I)^(-0.20))
di<- rocDI(DTC,t)
cx<- CX(DTC,t)
nk<- KM(DTC,t)

plot(km$FP,km$TP,type="s",lwd=2,xlab="1 - Especificidad",
ylab="Sensibilidad",cex.lab=1.2,xlim=c(0,1),ylim=c(0,1),cex.axis=1,col=2)
lines(c(0,1),c(0,1),col="gray",lty=2)
lines(ak$FP,ak$TP,type="s",lwd=2,col=3)
lines(di$FP,di$TP,type="s",lwd=2,col=4)
lines(1-cx$TN,cx$TP,type="s",lwd=2,col=7)
lines(1-nk$TN,nk$TP,type="s",lwd=2,col=6)
axis(2,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
axis(1,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
lines(c(0.7,0.8),c(0.5,0.5),lwd=2,col=7)
text(0.82,0.5,expression(N[C]),cex=0.8,pos=4)
lines(c(0.7,0.8),c(0.43,0.43),lwd=2,col=6)
text(0.82,0.43,expression(N[K]),cex=0.8,pos=4)
```

```
lines(c(0.7,0.8),c(0.36,0.36),lwd=2,col=2)
text(0.82,0.36,expression(K[M]),cex=0.8,pos=4)
lines(c(0.7,0.8),c(0.29,0.29),lwd=2,col=3)
text(0.82,0.29,expression(A[K]),cex=0.8,pos=4)
lines(c(0.7,0.8),c(0.22,0.22),lwd=2,col=4)
text(0.82,0.22,expression(D[I]),cex=0.8,pos=4)
if(t==4){text(0.2,0.9,"t=4",cex=1.4)}
if(t==7){text(0.2,0.9,"t=7",cex=1.4)}
if(t==10){text(0.2,0.9,"t=10",cex=1.4)}
if(t==13){text(0.2,0.9,"t=13",cex=1.4)}
}
```

Figura 5.3

```
library(survivalROC)
library(risksetROC)
library(KMsurv)
library(mvtnorm)
library(ks)
library(pROC)

#### COX

CX<- function(DT,t)
{
prob<- NULL
T<- DT[,1]; E<- DT[,2]; M<- DT[,3]
P<- which(T <= t & E==1)
N<- which(T > t)
IN<- which(T <= t & E==0)

if (length(IN)>0)
{
fit <- coxph(Surv(T, E) ~ M)
```

```
md<- survfit(fit, newdata=data.frame(DT))
prob<- 1:length(IN)
for (j in 1:length(IN))
{
f<- approxfun(c(0,md$time),c(1,md$surv[,IN[j]]))
prob[j]<- f(t)/f(T[IN[j]])
if (is.na(prob[j])) prob[j]<- 1
}
}
cut<- sort(c(min(as.numeric(names(table(M))))-1,as.numeric(names(table(M)))))
nS<- length(P) + sum(1-prob)
nE<- length(N) + sum(prob)
Se<- cut; Es<- cut
for (i in 1:length(cut))
{
Se[i]<- (sum(M[P]> cut[i]) + sum(1-prob[which(M[IN]>cut[i])]))/nS
Es[i]<- (sum(M[N]<= cut[i]) + sum(prob[which(M[IN]<=cut[i])]))/nE
}
list(TP=Se, TN=Es, prob=prob)
}

####  KAPLAN-MEIER

KM<- function(DT,t)
{
prob<- NULL
T<- DT[,1]; E<- DT[,2]; M<- DT[,3]
P<- which(T <= t & E==1)
N<- which(T > t)
IN<- which(T <= t & E==0)

if (length(IN)>0)
{
prob<- 1:length(IN)
for (j in 1:length(IN))
```

```
{
I<- which(M<= M[IN[j]])
fit<- survfit(Surv(T[I],E[I])~1)
f<- stepfun(fit$time,c(1,fit$surv))
prob[j]<- f(t)/f(T[IN[j]])
if (is.na(prob[j])) prob[j]<- 1
}
}
cut<- unique(sort(M))
nS<- length(P) + sum(1-prob)
nE<- length(N) + sum(prob)
Se<- cut; Es<- cut
for (i in 1:length(cut))
{
Se[i]<- (sum(M[P]> cut[i]) + sum(1-prob[which(M[IN]>cut[i])]))/nS
Es[i]<- (sum(M[N]<= cut[i]) + sum(prob[which(M[IN]<=cut[i])]))/nE
}
list(TP=Se, TN=Es, prob=prob)
}

##### ROC sin datos censurados

rocDI <- function(DT,t)
{
I1<- which(DT[,1]<= t & DT[,2]==1)
I0<- which(DT[,1]>t)
if (length(I1)*length(I0)==0) {FP<- c(0,0.5,1); TP<- c(0,0.5,1)}
else {
M<- c(DT[I0,3],DT[I1,3])
R<- c(rep(0,length(I0)),rep(1,length(I1)))
r<- roc(R,M)
FP<- 1-r$spec; TP<- r$sens}
list( FP= FP, TP=TP)
}
```

```
##### AUC EN EL TIEMPO

t<- c(0.5,1:15)
A<- t
AL<- t
AU<- t
p<- seq(0,1,0.01)
B<- 100
AB<- rep(0,B)

for (k in 1:16)
{
md<- CX(DTC,t[k])
cx<- approxfun(c(0,1-md$TN,1),c(0,md$TP,1))(p)
A[k]<- 0.01*max(sum(cx))

for(b in 1:B)
{
IB<- sample(1:(dim(DT)[1]),replace=TRUE)
I<- which(DT[IB,1]>=0 & DT[IB,2]>=0 & DT[IB,3]>=0)
DTB<- cbind(DT[IB[I],1],DT[IB[I],2],-DT[IB[I],3])
md<- CX(DTB,t[k])
cx<- approxfun(c(0,1-md$TN,1),c(0,md$TP,1))(p)
AB[b]<- 0.01*max(sum(cx))
}
AL[k]<- quantile(AB,0.025);
AU[k]<- quantile(AB,0.975);
print(k)
}

par(mar=c(5,5,1,1))
plot(t,A,type="b",lwd=5,xlab="Anos",
ylab="AUC",cex.lab=2.0,xlim=c(0,15),ylim=c(0.35,1.05),cex.axis=1.5)
lines(c(0,15),c(0.5,0.5),col="gray",lty=2)
lines(t,A,lwd=5)
```

```
lines(t,AL,lty=2)
lines(t,AU,lty=2)
axis(2,seq(0,1,0.1),lab=rep(" ",length(seq(0,1,0.1))))
axis(1,seq(0,15,1),lab=rep(" ",length(seq(0,15,1))))
```

Bibliografía

[1] Akritas M.G. Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*. 1994; **22**(3): 1299-1327.

[2] Burgueño M.J., García-Bastos J.L., González-Buitrago J.M. Las curvas ROC en la evaluación de las pruebas diagnósticas. *Med Clin*(Barc). 1995; **104**: 661-670.

[3] Cai T., Pepe M.S., Zheng Y., Lumley T., Jenny N.S. The sensitivity and specificity of markers for event times. *Biostatistics*. 2006; **7**(2): 182-197.

[4] Etzioni R., Pepe M.S., Longton G., Hu C., Goodman G. Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making*. 1999; **19**: 242-251.

[5] Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. 2006; **27**: 861-874.

[6] Gerds T.A., Michael M.W., Schumacher M., Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*. 2013; **32**(13): 2173-2184.

[7] Godoy A. Introducción al Análisis de Supervivencia con R (Tesis). Facultad de Ciencias, UNAM. 2009.

[8] Goncalves L., Subtil A., Oliveira M.R, Bermudez, P.Z. Roc Curve Estimation: An Overview. *Statistical Journal*. 2014; **12**(1): 1-20.

[9] Green D.M., Swets J.A. Signal detection theory and psychophysics. 1966. New York: Wiley.

[10] Heagerty P., Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005; **61**(1): 92-105.

[11] Heagerty P.J., Lumley T., Pepe M.S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; **56**: 337-344.

[12] Jacqmin-Gadda H., Blanche .P, Chary E., Touraine C., Dartigues J.F. Receiver operating characteristic curve estimation for time to event with semicompeting risks and interval censoring. *Statistical Methods in Medical Research*. 2014.

[13] Kaplan E.L., Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; **53**: 457- 481.

[14] Klein J.P., Moeschberger M.L. *Survival analysis: Techniques for censored and truncated data*. 2003. New York: Springer.

[15] Krzanowski W.J., Hand D.J. ROC curves for Continuous data. 2009. CRC Press, Chapman&Hall, Londres.

[16] Lee E.T. Statistical methods for survival data analysis. 1980. Lifetime Learning Publications, Belmont, CA.

[17] Lusted, L. Decision - making studies in patients management. *Journal of Medicine*. 1971; **284**: 416-424.

[18] Marin J.M., Alfageme I., Almagro P., Casanova C., Esteban C., Soler-Cataluña J.J., de Torres J.P., Martínez-Camblor P., Miravitlles M., Celli B.R., Soriano J. Multicomponent indices to predict mortality in COPD: the COCOMICS study. *European Respiratory Journal*. 2013; **42**: 323-332.

[19] Martinez E. Z., Louzada-Neto F., Pereira B. B. A curva ROC para testes diagnósticos. it Cadernos Saúde Coletiva. 2003; **11**: 7-31.

[20] Martínez-Camblor P. Area under the ROC curve comparison in the presence of missing data. *Journal of the Korean Statistical Society*. 2013; **42**(4): 431-442.

[21] Martínez-Camblor P. Nonparametric cutoff point estimation for diagnostic decisions with weighted errors. *Revista Colombiana de Estadística*. 2011; **34**(1): 133-146.

[22] Martínez-Camblor P., Corral N., Rey C., Pascual J., Cernuda-Morollón E. ROC curve generalization for non-monotone relationships. *Statistical Methods in Medical Research*. 2014.

[23] Martínez-González M.A., Alonso A., Fidalgo J.L. ¿Qué es una hazard ratio? Nociones de análisis de supervivencia. *M*edicina Cliníca (Barcelona). 2008; **131**: 65-72.

[24] Martínez-Camblor P. Comparación de pruebas diagnósticas desde la curva ROC. *Revista Colombiana de Estadística*. 2007; **30**(2): 163-176.

[25] Park S.H., Goo J.M., Jo C.H. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean Journal of Radiology*. 2004; **5**(1): 11-18.

[26] Pepe M.S. The Statistical Evaluation of Medical Tests for Classification and Prediction. 2003. Oxford, Oxford University Press.

[27] Rupert G. Miller, Jr. Survival Analysis. 1975. John Wiley and Son, New York.

[28] Salanti G., Kurt U. A non-parametric framework for estimating threshold limit values. *BMC Medical Research Methodology*. 2005; **5**: 36.

[29] Slate E., Turnbull B. Statistical models for longitudinal biomarker of disease onset. *Statistics in Medicine*. 2000; **19**: 617-637.

[30] Soriano J., Alfageme I., Almagro P., Casanova C., Esteban C., Soler-Cataluña J.J., de Torres J.P., Martínez-Camblor P., Miravitlles M., Celli B.R., Marin J.R. Distribution and prognostic validity of the new global initiative for chronic obstructive lung disease grading classification. *Chest*. 2013; **143**(3): 694-702.

[31] Spackman K.A. Signal detection theory: Valuable tools for evaluating inductive learning. 1989. Proc. Sixth Internat. Workshop on Machine Learning. Morgan Kaufman, San Mateo, CA.

[32] Swets J.A., Dawes R.M., Monahan J. Better decisions through science. *Scientific American*. 2000; **283**: 82-87.

[33] Torres A. Curvas ROC para Datos de Supervivencia. Aplicación a Datos Biomédicos. 2010. Proyecto presentado para culminar el Máster en Técnicas Estadísticas de la Universidad de Santiago de Compostela.

[34] Venkatraman E.S., Begg C.B. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*. 1996; **83**(4): 835-848.

[35] Viallon V., Latouche A. Discrimination measures for survival outcomes: Connection between the AUC and the predictiveness curve. *Biometrical Journal*. 2011; **53**: 217-236.

[36] Walters S.J. What is a Cox model? *Aventis [serie online]*. 2003; **1**(10), disponible en: `www.evidence-based-medicene.co.uk.data`

[37] Wolf P., Schmidt G., Ulm K. The use of ROC for defining the validity of the prognostic index in censored data. *Statistics and Probability Letters*. 2011; **81**: 783-791.

[38] Yousef W.A., Kundu S., Wagner R.F. Nonparametric estimation of the threshold at an operating point on the ROC curve. *Computational Statistics and Data Analysis*. 2009; **33**(12): 4370-4383.

[39] Zhou, X.H., Obuchowski N.A., McClich D.K. Statistical methods in diagnostic medicine. 2011. Second Edition, John Wiley & Sons, New Jersey.