Introducción a las curvas *ROC* para clasificadores binarios

Jorge Palacio Lacasta

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Director del trabajo: José Tomás Alcalá Nalvaiz
28 de Junio de 2021

Prólogo

La curva *ROC* (receiver operating characteristic) es una representación gráfica la cuál mide la capacidad diagnóstica de un clasificador binario al variar su umbral de decisión. Su nacimiento surge durante la Segunda Guerra Mundial en el contexto del análisis de señales de radar. El objetivo era medir la eficacia para detectar objetos enemigos en el campo de batalla mediante señales de radar.

Dicha curva se puede representar como una expresión directa entre funciones de distribución de variables aleatorias por lo que se pueden aplicar técnicas estadísticas para estimarla. Las propiedades matemáticas no se han estudiado con gran profundidad hasta épocas relativamente recientes.

El análisis *ROC* se utiliza de forma muy extensa en epidemiología o en radiología, de tal modo que se encuentra muy relacionado con la medicina basada en la evidencia. Más recientemente dichas curvas han sido muy utilizadas a la hora de evaluar clasificadores binarios en el contexto del aprendizaje automático supervisado.

En este trabajo se introduce al lector en los principales conceptos relacionados con las curvas *ROC* y sus diversas formas de estimación. En la parte final se hace hincapié en cómo modelar la influencia de las covariables en las curvas, mostrando una aplicación a datos reales como caso práctico.

Abstract

The *ROC* curve is one of the most important graphical tools for evaluating the diagnostic power of a binary classifier. Its use is ubiquitous in various applied statistical fields. The objective of this work is to review the main mathematical properties of the population curve, as well as to describe the main statistical methods to estimate it. Only the top ideas related to these methods will be displayed. Finally, we consider the presence of a covariate related to the score variable and analyze how to incorporate this additional information in the *ROC* curves. We ended up applying this methodology to a real data set using the latest R software libraries related to this topic.

The content is divided in 3 chapters. In chapter 1, the *ROC* curve is defined as a curve that depends on distribution functions and its main geometric properties are studied. Subsequently, the concept of classifier is defined, although the focus is on the study of binary classifiers. The hypothesis tests and diagnostic tests as methods used for binary classification are introduced. The concepts of sensitivity and specificity are mentioned as statistical indicators to evaluate the degree of efficiency of a diagnostic test. Then, the most common indices for the curve like the area under the curve (AUC) or the Youden index are introduced. To conclude the chapter, formulas for the *ROC* curve and the corresponding indices are proposed, assuming that our random variables have a normal distribution.

In chapter 2, different procedures of estimating the curve and confidence intervals for these estimates are shown. First, a direct empirical estimation is proposed and is based on empirical distribution functions. Then, a parametric estimation is displayed in which the binormal method is studied in detail. After that, an estimation and confidence intervals of the AUC are proposed. Non-parametric estimation methods are also considered in which kernel functions are important for estimating density functions. Finally, hypothesis test applied to *ROC* curves in which the contrast statistic depends on the estimated AUC values are proposed. For this chapter we have mainly made use of [8].

In chapter 3, the concept of covariate-dependent *ROC* curve is introduced. It is important to take into account the covariates, since the information provided by them can modify the *ROC* curve. Thus, *ROC* curves conditioned to a value of a covariate or covariate-adjusted *ROC* curve can be obtained. As in the previous chapter, several methods of estimating the conditional curve are also proposed here. In this section, we have based our study on [10]. To conclude the chapter, a real case is displayed whose objective is to try to relate the Body Mass Index with the detection of patients at cardiovascular risk. For that, we propose a first estimation of the non-covariate-dependent curve and we study whether sex (a discrete covariate) can influence the curve. As sex has an influence, men and women data are studied separately. Also, we add the age as a continuous covariate and we estimate the curve using the methods proposed at the beginning of this chapter.

In addition, the R code proposed to perform the case study can be consulted in the appendix.

Índice general

Capítulo 1

Curva *ROC* poblacional

Tras varias definiciones equivalentes de la curva *ROC*, analizamos sus principales propiedades geométricas, alguna de ellas relacionada con el orden estocástico de las variables aleatorias que definen la curva *ROC*. Además, tras introducir los conceptos de sensibilidad y especifidad, redefinimos la curva *ROC* en función de dichos conceptos. Para finalizar el capítulo, describimos brevemente algunos de los índices respecto la curva y se proponen expresiones explícitas para el modelo poblacional binormal.

1.1. Definición y primeras propiedades

Sean X e Y dos variables aleatorias absolutamente continuas con F y G sus funciones de distribución. Asumiremos, salvo que se especifique lo contrario, que F y G son derivables con respectivas funciones de densidad f y g.

Definición 1. *La curva ROC en su forma paramétrica se define como la función*

$$\begin{array}{rccl} ROC\colon & \mathbb{R} & \longrightarrow & [0,1]\times[0,1] \\ & c & \longmapsto & ROC(c)=(1-F(c),1-G(c)). \end{array} \tag{1.1}$$

Proposición 1.1. *Los puntos* $(0,0)$ *y* $(1,1)$ *pertenecen a la curva ROC.*

Demostración. Utilizando la forma paramétrica de la curva definida en (1.1) y que F y G son funciones de distribución, tenemos que: $\lim_{c\to-\infty} 1-F(c)=\lim_{c\to-\infty} 1-G(c)=1\Rightarrow \lim_{c\to+\infty} ROC(c)=(1,1)$ y $\lim_{c\to+\infty} 1-F(c)=\lim_{c\to+\infty} 1-G(c)=0\Rightarrow \lim_{c\to+\infty} ROC(c)=(0,0)$. □

Proposición 1.2. *(Adaptada de [2]) Utilizando la forma explícita de una curva, podemos definir la curva ROC como*

$$\begin{array}{rccl} ROC\colon & [0,1] & \longrightarrow & [0,1] \\ & t & \longmapsto & ROC(t)=1-G(F^{-1}(1-t)), \end{array} \tag{1.2}$$

donde $F^{-1}(t)=\text{ínf}_{u\in\mathbb{R}}\{\ u\ t.q.F(u)\geq t\}$.

Demostración. Partiendo de la curva *ROC* dada en (1.1) e identificando $t=1-F(c)$ entonces $c=F^{-1}(1-t)$ por ser X absolutamente continua. Ahora sustuimos en la segunda componente de la curva y tenemos $1-G(c)=1-G(F^{-1}(1-t))$. □

Observación 1. *Si consideramos* $\bar{F}=1-F$ *y* $\bar{G}=1-G$ *entonces, podemos definir* $ROC(t)=\bar{G}\circ(\bar{F}^{-1}(t))$, $\forall t\in[0,1]$ *aplicando la defínción de* F^{-1}.

A continuación vamos a demostrar una serie de propiedades de la curva relativas a su forma. Comenzamos analizando la invarianza y monotonía de este tipo de curvas.

Proposición 1.3. *(Adaptada de [2]) La curva ROC es invariante frente a transformaciones monótonas estrictamente crecientes.*

Demostración. Sea h la transformación estrictamente creciente y consideramos $W_1 = h(X)$ y $W_2 = h(Y)$. Sea $(\bar{F}(c), \bar{G}(c))$ el punto de la curva *ROC* en el punto c para X e Y y denotamos $d = h(c)$. Entonces se deduce que $\bar{F}(c) = 1 - F(c) = P(X > c) = P(W_1 > d)$ y que $\bar{G}(c) = 1 - G(c) = P(Y > c) = P(W_2 > d)$. Por tanto el mismo punto se encuentra en la curva ROC para W_1 y W_2.

□

Para dar una demostración simple de la propiedad de que la curva *ROC* es no decreciente nos centramos en la situación regular. En este sentido, asumimos que F y G poseen densidades de Lebesgue continuas, estrictamente positivas, en el interior de su soporte común.

Proposición 1.4. *(Adaptada de [2]) La curva ROC es una función monótona creciente en* $[0,1]$ *con pendiente*

$$\frac{dROC(t)}{dt} = \frac{g(\bar{F}^{-1}(t))}{f(\bar{F}^{-1}(t))} > 0, \quad 0 < t < 1. \tag{1.3}$$

Demostración. Considerando la forma explícita de la curva definida en (1.2) , teniendo en cuenta que $\frac{d\bar{G}(s)}{ds} = -g(s)$ y tomando $w = \bar{F}^{-1}(t)$, entonces

$$\frac{d\bar{F}^{-1}(t)}{dt} = \frac{1}{\frac{d}{dw}\bar{F}(w)} = \frac{1}{-f(w)} = \frac{1}{-f(\bar{F}^{-1}(t))}. \tag{1.4}$$

Por lo tanto, para $0 < t < 1$, obtenemos que:

$$\frac{dROC(t)}{dt} = \frac{d\bar{G}(\bar{F}^{-1}(t))}{dt} = -g(\bar{F}^{-1}(t))\frac{d\bar{F}^{-1}(t)}{dt} = \frac{-g(\bar{F}^{-1}(t))}{-f(\bar{F}^{-1}(t))} = \frac{g(\bar{F}^{-1}(t))}{f(\bar{F}^{-1}(t))}.$$

Como las funciones de densidad g y f son siempre estrictamente positivas, entonces la derivada de la curva *ROC* es mayor que 0. □

La siguiente propiedad geométrica de la curva *ROC* está directamente relacionada con la ordenación de las variables X e Y en un sentido estocástico. La definición que se muestra a continuación se puede encontrar en [2].

Definición 2. *Sean X e Y variables aleatorias. Decimos que X es más pequeña que Y en cociente de verosimilitud, y se expresa como $X \leq_{lr} Y$, si $\frac{g(x)}{f(x)}$ crece cuando x se encuentra en la unión de los soportes de X e Y. Equivalentemente, si*

$$f(x)g(y) \geq f(y)g(x) \quad \forall x < y.$$

Proposición 1.5. *(Adaptada de [2]) Sean X e Y variables aleatorias absolutamente continuas con sus respectivas funciones de densidad derivables, entonces $X \leq_{lr} Y \Leftrightarrow ROC(t)$ es cóncava.*

Demostración. $\Rightarrow$) Por la Proposición 1.4 tenemos que

$$\frac{d\bar{G}(\bar{F}^{-1}(t))}{dt} = \frac{g(\bar{F}^{-1}(t))}{f(\bar{F}^{-1}(t))}.$$

Si ahora volvemos a derivar respecto de t obtenemos

$$\frac{d^2\bar{G}(\bar{F}^{-1}(t))}{d^2t} = \frac{d\frac{g(\bar{F}^{-1}(t))}{f(\bar{F}^{-1}(t))}}{dt} = \frac{d\frac{g(\bar{F}^{-1}(t))}{f(\bar{F}^{-1}(t))}}{d\bar{F}^{-1}(t)}\frac{d\bar{F}^{-1}(t)}{dt}.$$

Usando la Definición 2 y como las funciones de densidad son derivables, tenemos que $\frac{d\frac{g(t)}{f(t)}}{dt} \geq 0$ y además sabemos por (1.4) que $\frac{d\bar{F}^{-1}(t)}{dt} = \frac{-1}{f(\bar{F}^{-1}(t))} < 0$ al ser f estrictamente positiva. Luego

$$\frac{d^2\bar{G}(\bar{F}^{-1}(t))}{dt} \leq 0.$$

$\Leftarrow$) Si $ROC(t)$ es cóncava entonces $\frac{d^2\bar{G}(\bar{F}^{-1}(t))}{dt} \leq 0$, además sabemos que $\frac{d^2\bar{G}(\bar{F}^{-1}(t))}{dt} = \frac{d\frac{g(\bar{F}^{-1}(t))}{f(\bar{F}^{-1}(t))}}{d\bar{F}^{-1}(t)} \frac{d\bar{F}^{-1}(t)}{dt}$ y como por (1.4) tenemos que $\frac{d\bar{F}^{-1}(t)}{dt} < 0$ entonces $\frac{d\frac{g(t)}{f(t)}}{dt} \geq 0$. □

Corolario 1.1. *Bajo las condiciones de la proposición anterior, si $X \leq_{lr} Y$, entonces la curva ROC se encuentra por encima de la diagonal principal.*

Demostración. Considerando el resultado de las Proposiciones 1.1 y 1.5 entonces la curva *ROC* es cóncava con $ROC(0) = 0$ y $ROC(1) = 1$. □

1.2. Clasificadores binarios y curva *ROC*

La clasificación de objetos es su asignación a uno de los posibles grupos definidos. Los objetos se definen por sus características, por lo que el problema se reduce a definir unas fronteras entre clases o equivalentemente a definir unas reglas de asignación.

Definición 3. *La clasificación estadística, como parte de los métodos de aprendizaje automático consiste en estimar o construir estas reglas en base a un conjunto de entrenamiento de casos previamente etiquetados. Por ejemplo, en el caso binario mediante una variable de decisión dicotómica.*

1.2.1. Test de hipótesis

Recordamos ahora el concepto del contraste de hipótesis en estadistica ya que podemos plantearlo como una decisión o una clasificación binaria a través de un test o un estadistico de contraste.

Definición 4. *Una hipótesis estadística es una afirmación o proposición respecto a alguna característica de una población. Contrastar una hipótesis es comparar las predicciones con la realidad que observamos mediante una muestra.*

Consideramos la hipótesis nula como H_0, la hipótesis alternativa como H_1 y la región de rechazo como el conjunto de los valores del test para el cual H_0 es rechazada, ya que al comienzo del estudio siempre se elige H_0 para que sea la hipótesis que se asuma como válida.

Por lo tanto, podemos considerar un test de hipótesis como una forma de evaluar la evidencia que los datos proporcionan para probar la hipótesis nula (H_0) y así clasificar binariamente a los individuos.

Cuando realizamos un test de hipótesis, podemos cometer dos tipos de errores:

1. Error de tipo *I*. Se comete cuando la hipótesis nula es verdarera pero es rechazada. La probabilidad de que esto ocurra se denota por α.

2. Error de tipo *II*. Se comete cuando la hipótesis nula debería ser rechazada pero se acepta y denotamos la probabilidad de que esto ocurra como β.

1.2.2. Pruebas diagnósticas

Es frecuente disponer de una variable medible sobre los pacientes que puede determinar si se posee o no una característica o condición concreta (supuestamente patológica) que no es observable de forma directa. La prueba diagnóstica es una variable de decisión basada en la medición realizada sobre un paciente que lo clasifica como positivo (presencia de la característica) o negativo (ausencia de característica). De forma tácita, entendemos que a mayor valor en la medición mayor posibilidad de que la característica buscada esté presente.

Ahora definimos la variable aleatoria $Z \backsim Be(p)$ con $p \in [0,1]$. Al parámetro p lo llamaremos prevalencia del evento sobre la población y representa la probabilidad de presentar el evento. Así pues,

$$Z = \begin{cases} 0, & \text{si el individuo no presenta el evento} \\ 1, & \text{si el individuo presenta el evento.} \end{cases}$$

Consideramos ahora la variable aleatoria real T que mide ciertas caracteristicas de los individuos y podemos definir la variable aleatoria D (variable de decisión) como una función indicadora que depende de T,

$$D = \mathbb{1}_{T>c} = \begin{cases} 1, & \text{si } T > c \\ 0, & \text{si } T \leq c \end{cases}$$

donde el punto c lo denominamos valor umbral o punto de corte. De tal forma que $E(\mathbb{1}_{T>c}) = P(T > c)$.

Definición 5. *Se define sensibilidad (la denotamos con S) como la probabilidad de, dado un individuo que presenta un cierto evento o característica, clasificarlo correctamente. También se define como probabilidad de verdadero positivo.*

$$S = P(D = 1|Z = 1) = \frac{P(D = 1 \cap Z = 1)}{P(Z = 1)} = \frac{P(T > c \cap Z = 1)}{P(Z = 1)}.$$

Definición 6. *Se define especificidad (E) como la probabilidad de, dado un individuo que no presenta una cierta característica, que la prueba lo clasifique correctamente. También se define como probabilidad de verdadero negativo.*

$$E = P(D = 0|Z = 0) = \frac{P(D = 0 \cap Z = 0)}{P(Z = 0)} = \frac{P(T \leq c \cap Z = 0)}{P(Z = 0)}.$$

Definición 7. *Definimos probabilidad de falso positivo como la probabilidad de error de tipo I. Dicha expresión viene dada por:*

$$P(D = 1|Z = 0) = \frac{P(D = 1 \cap Z = 0)}{P(Z = 0)} = \frac{P(T > c \cap Z = 0)}{P(Z = 0)} = \alpha = 1 - E.$$

Definición 8. *Definimos probabilidad de falso negativo como la probabilidad de error de tipo II.*

$$P(D = 0|Z = 1) = \frac{P(D = 0 \cap Z = 1)}{P(Z = 1)} = \frac{P(T \leq c \cap Z = 1)}{P(Z = 1)} = \beta = 1 - S.$$

	$Z=1$	$Z=0$
$D=1$	S	$1-E$
$D=0$	$1-S$	E

Cuadro 1.1: Probabilidades condicionadas de las variables decisoras (D), dada la variable de estado (Z).

Un ejemplo de prueba dignóstica puede ser una prueba PCR, que nos clasifica a los individuos como positivos o negativos en COVID-19. Si definimos el CT (Cycle Threshold) como el número de ciclos de amplificación de la carga vírica, el test se considera positivo si el CT es menor que 30, véase [4].

En el Cuadro 1.2 se muestra la relación entre los test de hipósteis y las pruebas diagnósticas.

	Test de Hipótesis	Prueba Diagnóstica
Posibles estados	H_0 vs H_1	$Z=0$ vs $Z=1$
Información	Datos para n sujetos	Resultado de un test para un sujeto
Error de tipo I	α	$1-E$
Error de tipo II	β	$1-S$

Cuadro 1.2: Analogía entre los test de hipótesis y las pruebas diagnósticas.

1.2.3. Sensibilidad y especificidad en la curva *ROC*

Ahora consideramos $X = T|Z=0$ e $Y = T|Z=1$ donde sus respectivas funciones de distribución son $F(c) = P(X \leq c)$ y $G(c) = P(Y \leq c)$.

Proposición 1.6. *(Adaptada de [15]) La curva ROC en forma paramétrica se puede expresar como*

$$ROC(c) = \{(1-E(c), S(c)),\ c \in (-\infty, \infty)\} \tag{1.5}$$

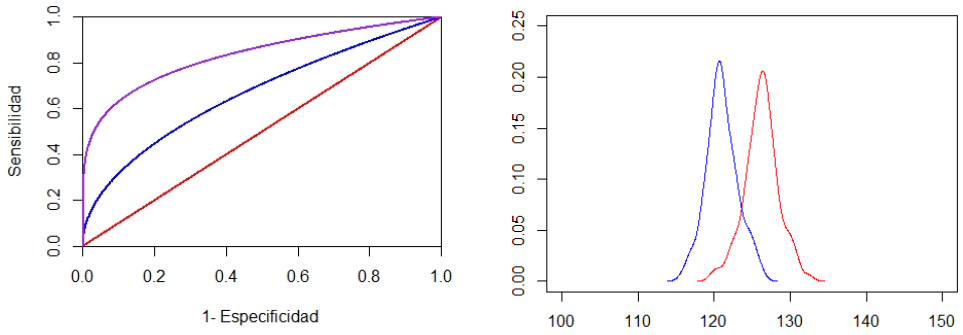
Demostración. Dado un c fijo, recordemos que $D = \mathbb{1}_{T>c}$. Tenemos que los verdaderos positivos generados por c pueden representarse como:

$$S = P(D=1|Z=1) \Rightarrow S(c) = P(T>c|Z=1) = P(Y>c) = 1-P(Y \leq c) = 1-G(c) \tag{1.6}$$

y los falsos positivos como:

$$1-E = P(D=1|Z=0) \Rightarrow 1-E(c) = P(T>c|Z=0) = P(X>c) = 1-P(X \leq c) = 1-F(c). \tag{1.7}$$

□



(a) Ejemplo de curvas *ROC* (b) Funciones de densidad para ambos grupos

Figura 1.1: Relación entre curvas *ROC* y funciones de densidad

En la Figura 1.1a, se enfrenta la probabilidad de obtener verdaderos positivos (eje y) frente a la probabilidad de obtener falsos positivos (eje x). Cabe destacar la forma cóncava que presenta la curva *ROC*.

Cualquier curva que aparezca en el triángulo inferior derecho tiene un comportamiento peor que la clasificación al azar (la recta $y = x$). En ese caso, si se intercambian los resultados del clasificador, sus verdaderos positivos se convierten en falsos negativos y sus falsos positivos se convierten en verdaderos negativos. Así, siempre que haya una curva en el triángulo inferior derecho puede invertirse para producir una curva en el triángulo superior izquierdo.

Si las funciones de densidad dadas en 1.1b estuviesen muy solapadas entonces la capacidad discriminante de la prueba sería muy baja y por tanto, nuestro caso óptimo sería cuando no existiera solapamiento.

1.3. Índices de la curva *ROC*

Hemos visto anteriormente que la curva *ROC* es un resumen conveniente del conjunto de información que se necesitaría para una descripción del rendimiento de un clasificador sobre todos sus posibles valores umbrales. Sin embargo, incluso un resumen de este tipo puede ser demasiado complicado en algunas circunstancias, por lo que el interés se ha centrado en derivar resúmenes más simples. Varios índices de este tipo son de uso común, y ahora vamos a presentar los más populares.

1.3.1. Área bajo la curva

Uno de los índices numéricos más importantes para medir la capacidad discriminatioria de una prueba diagnóstica es el área bajo la curva.

Definición 9. *Se define el área bajo la curva como:*

$$AUC = \int_0^1 ROC(t)dt. \tag{1.8}$$

Proposición 1.7. *Si $X \leq_{lr} Y$, la curva ROC satisface que* $0.5 \leq AUC \leq 1$.

Demostración. Sabemos por el Corolario 1.1 que la curva *ROC* siempre se situa por encima de la diagonal $y = x$. Además, es una porción de área de un cuadrado de lado 1. □

Proposición 1.8. *Sean ROC_A y ROC_B dos curvas ROC distintas. Si $ROC_A \geq ROC_B$, es decir, si ROC_A no se encuentra en ningún lugar por debajo de ROC_B, entonces $AUC_A \geq AUC_B$.*

Demostración. Aplicando las propiedades de las integrales tenemos que si $ROC_A(t) \geq ROC_B(t) \Rightarrow \int_0^1 ROC_A(t)dt \geq \int_0^1 ROC_B(t)dt$. □

Cabe destacar que el recíproco no es cierto debido a la posibilidad de que las dos curvas puedan cruzarse entre sí.

Proposición 1.9. *(Adaptada de [8, p.27]) El área bajo la curva se puede expresar como:*

$$AUC = P(Y > X). \tag{1.9}$$

Demostración. Como $ROC(t) = 1 - G(c)$ y $t = 1 - F(c) \Rightarrow \frac{dt}{dc} = -F'(c) = -f(c)$. Teniendo en cuenta que cuando $c \to \infty \Rightarrow t \to 0$ y $c \to -\infty \Rightarrow t \to 1$, si planteamos un cambio en la variable de integración, tenemos:

$$\begin{aligned} AUC &= \int_0^1 ROC(t)dt = \int_\infty^{-\infty} (1 - G(c))\frac{dt}{dc}dc = \int_{-\infty}^{\infty} (1 - G(c))f(c)dc = \\ &= \int_{-\infty}^{\infty} P(Y > X|X = c)f(c)dc = P(Y > X). \end{aligned}$$

Aplicando en el último paso el Teorema de la probabilidad total. □

La escala de Swets, como se nombra en [15], interpreta los diferentes valores del *AUC*. Considera que tienen baja exactitud si el $AUC = [0,5; 0,7)$ que los valores son útiles para algunos propósitos si el $AUC = [0,7; 0,9)$ y que tienen una alta exactitud alta si el $AUC = [0,9; 1]$.

1.3.2. Área parcial bajo la curva

Hay en ocasiones en las que nos puede resultar interesante una probabilidad especifica de verdaderos positivos, por ejemplo, puede ser que en el punto t_0 entonces $ROC(t_0)$ nos proporciona un valor que nos interesa. También en otras ocasiones nos puede interesar el área solo en un rango de valores.

Definición 10. *Sea $0 < a < b < 1$, se define el área parcial bajo la curva en el intervalo (a,b) como*

$$PAUC(a,b) = \int_a^b ROC(t)dt. \tag{1.10}$$

1.3.3. Índice de Youden

Definición 11. *Definimos el índice de Youden como la máxima diferencia entre la probabilidad de verdadero positivo y de falso positivo, es decir,*

$$YI = \max_c(S(c) + E(c) - 1). \tag{1.11}$$

Este índice toma valores en el intervalo $[0,1]$ siendo los valores cercanos a 0 los correspondientes a una prueba con una capacidad discriminatoria prácticamente nula y los cercanos a 1 a una prueba perfecta ya que tendríamos valores de la sensiblidad y especificidad cercanos a 1.

Proposición 1.10. *Para variables aleatorias X e Y tales que $X \leq_{lr} Y$ y asumiendo la derivabilidad de sus respectivas funciones de densidad, el punto c para el cual se alcanza el mayor valor para el índice de Youden es el punto que satisface que $\frac{g(c)}{f(c)} = 1$.*

Demostración. Sea $\psi(c) = S(c) + E(c) - 1 = F(c) - G(c)$. Derivando $\psi(c)$ respecto a c e igualando a 0 obtenemos:

$$\frac{d(\psi(c))}{dc} = f(c) - g(c) = 0 \Rightarrow f(c) = g(c).$$

Volviendo a derivar respecto a c tenemos que

$$\frac{d^2\psi(c)}{d^2c} = f'(c) - g'(c).$$

Pero aplicando la Definición 2 y como hemos asumido derivabilidad de f y g, se deduce que $\frac{d\frac{g(t)}{f(t)}}{dt} \geq 0$. Derivando esta expresión y teniendo en cuenta que $f(c) = g(c)$ entonces $f'(c) - g'(c) \leq 0$. Luego en el punto c se alcanza un máximo. □

1.3.4. Tasa de verosimulitud

La tasa de verosimilitud es el cociente de la probabilidad de tener una respuesta positiva o negativa, bajo presencia de una cierta característica entre la probabilidad de respuesta positiva o negativa, bajo ausencia de la característica. Se utiliza para evaluar la calidad de una prueba diagnóstica, no depende de la prevalencia y se puede definir en función de la sensibilidad y especificidad. Las definiciones que se presentan a continación se pueden ver en [15, *p*.28].

Definición 12. *La tasa de verosimilitud positiva (LPR) viene dada por*

$$LRP = \frac{P(D=1|Z=1)}{P(D=1|Z=0)} = \frac{S}{1-E}, \tag{1.12}$$

es decir, la calcularemos cuando obtengamos una prueba sea positiva.

Definición 13. *La tasa de verosimilitud negativa (LNR) se define como*

$$LNP = \frac{P(D=0|Z=1)}{P(D=0|Z=0)} = \frac{1-S}{E} \tag{1.13}$$

y se calcula cuando el resultado de la prueba sea negativo.

Es claro que dichas tasas toman valores en el intervalo $[0,\infty)$.

1.3.5. Valor predictivo

El valor predictivo mide la eficacia real de una prueba diagnóstica y da la probabilidad de padecer o no una característica una vez conocido el resultado de la prueba diagnóstica. Se trata de un valor post-test y depende de la prevalencia.

Definición 14. *El valor predictivo positivo es la probabilidad de tener una cierta característica si el resultado de la prueba diagnóstica es positivo.*

$$VPP = P(Z = 1|D = 1) = \frac{P(Z = 1 \cap D = 1)}{P(D = 1)} \tag{1.14}$$

Definición 15. *Se define valor predictivo negativo como la probabilidad de no tener una cierta característica si el resultado de la prueba diagnóstica es negativo.*

$$VPN = P(Z = 0|D = 0) = \frac{P(Z = 0 \cap D = 0)}{P(D = 0)}. \tag{1.15}$$

Proposición 1.11. *Sean VPP y VPN definidas como en (1.14) y (1.15), podemos obtener las siguientes expresiones:*

$$VPP = \frac{pS}{pS + (1-p)(1-E)} \qquad VPN = \frac{(1-p)E}{(1-p)E + p(1-S)}. \tag{1.16}$$

Demostración.

$$\begin{aligned} VPP &= \frac{P(Z = 1 \cap D = 1)}{P(D = 1)} = \frac{P(Z = 1 \cap D = 1)}{P(Z = 1)P(D = 1|Z = 1) + P(Z = 0)P(D = 1|Z = 0)} = \\ &= \frac{P(Z = 1)P(D = 1|Z = 1)}{P(Z = 1)P(D = 1|Z = 1) + P(Z = 0)P(D = 1|Z = 0)} = \frac{pS}{pS + (1-p)(1-E)}. \end{aligned}$$

De manera análoga se obtiene $VPN = \frac{(1-p)E}{(1-p)E+p(1-S)}$. □

1.4. Modelo binormal

Para finalizar este primer capítulo, veamos cómo es la curva cuando ambas variables aleatorias siguen una distribución normal y obtendremos una forma explícita de alguno de los índices de la curva. En [15], se citan varios trabajos en los que se comprueba que el modelo binormal es un modelo robusto en el sentido de que aunque cierta cantidad de observaciones no siguen una distribución normal, ofrece buenos resultados.

Suponemos que las variables $Y \backsim N(\mu_y, \sigma_y)$ y $X \backsim N(\mu_x, \sigma_x)$, que $\mu_y \geq \mu_x$ y $\sigma_x, \sigma_y > 0$. Es claro que si denotamos $Z_y = \frac{Y-\mu_y}{\sigma_y}$, $Z_x = \frac{Y-\mu_x}{\sigma_x} \Rightarrow Z_y \backsim N(0,1)$ y $Z_x \backsim N(0,1)$. Como sabemos por (1.6) y (1.7) que $S(c) = P(Y > c)$ y $1 - E(c) = P(X > c)$, tenemos:

$$1 - E(c) = P(X > c) = P\left(Z_x > \frac{c-\mu_x}{\sigma_x}\right) = 1 - P\left(Z_x \leq \frac{c-\mu_x}{\sigma_x}\right) = 1 - \Phi\left(\frac{c-\mu_x}{\sigma_x}\right) = \Phi\left(\frac{\mu_x - c}{\sigma_x}\right)$$

donde Φ representa la función de distribución de la normal estándar. Si denotamos ahora $z_x = \Phi^{-1}(1 - E) = \frac{\mu_x - c}{\sigma_x}$, entonces $c = \mu_x - \sigma_x z_x$. De modo análogo, teniendo en cuenta la distribución de la variable aleatoria Y:

$$S(c) = P(Y > c) = P\left(Z_y > \frac{c-\mu_y}{\sigma_y}\right) = 1 - P\left(Z_y \leq \frac{c-\mu_y}{\sigma_y}\right) = 1 - \Phi\left(\frac{c-\mu_y}{\sigma_y}\right) = \Phi\left(\frac{\mu_y - c}{\sigma_y}\right).$$

Sustituyendo el valor de c en esta ecuación, obtenemos que $S = \Phi\left(\frac{\mu_y - (\mu_x - \sigma_x \cdot z_x)}{\sigma_y}\right)$.

Entonces la curva *ROC* tiene la forma

$$(1-E,S) = (\Phi(z_x), \Phi(a+b\cdot z_x)) \tag{1.17}$$

donde $a = \frac{\mu_y-\mu_x}{\sigma_y}$ y $b = \frac{\sigma_x}{\sigma_y}$. El valor de a nos da el intercepto de la curva con $a \geq 0$ y b nos da el valor de la pendiente de la curva cuyo valor es no negativo.

Podemos ahora encontrar una expresión sencilla y simplificada del AUC. Hemos visto en (1.9) que $AUC = P(Y > X) = P(Y - X > 0)$. Si volvemos a asumir como antes que $Y \backsim N(\mu_y, \sigma_y)$ y $X \backsim N(\mu_x, \sigma_x)$, entonces $Y - X \backsim N(\mu_y - \mu_x, \sqrt{\sigma_y^2 + \sigma_x^2})$ aplicando la definición de varianza, esperanza y asumiendo independencia. Si consideramos

$$Z = \frac{(Y-X)-(\mu_y-\mu_x)}{\sqrt{\sigma_y^2+\sigma_x^2}} \backsim N(0,1)$$

entonces tenemos

$$\begin{aligned} AUC &= P(Y-X>0) = P(Z > 0 - \frac{(\mu_y-\mu_x)}{\sqrt{\sigma_y^2+\sigma_x^2}}) = 1 - P(Z \leq -\frac{(\mu_y-\mu_x)}{\sqrt{\sigma_y^2+\sigma_x^2}}) = \\ &= 1 - \Phi(\frac{-(\mu_y-\mu_x)}{\sqrt{\sigma_y^2+\sigma_x^2}}) = \Phi(\frac{\mu_y-\mu_x}{\sqrt{\sigma_y^2+\sigma_x^2}}) = \Phi(\frac{a}{\sqrt{1+b^2}}). \end{aligned} \tag{1.18}$$

No existe una forma analítica sencilla para $PAUC(a,b)$, aunque se puede evaluar mediante integración numérica.

Para el índice de Youden definido en (1.11), podemos obtener la expresión

$$YI = \max_c(S(c)+E(c)-1) = \max_c(\Phi(\frac{\mu_y-c}{\sigma_y}) - \Phi(\frac{\mu_x-c}{\sigma_x})). \tag{1.19}$$

Como sabemos por la Proposición 1.10, el punto c es el que cumple que $\frac{g(c)}{f(c)} = 1$, si despejamos c, obtenemos el valor

$$c = \frac{\mu_y\sigma_x^2 - \mu_x\sigma_y^2 - \sigma_x\sigma_y\sqrt{(\mu_y-\mu_x)^2 + (\sigma_x^2-\sigma_y^2)log(\frac{\sigma_x^2}{\sigma_y^2})}}{\sigma_x^2-\sigma_y^2} \tag{1.20}$$

y si se asume que $\sigma_x = \sigma_y$, el valor $c = \frac{\mu_x+\mu_y}{2}$.

Para la tasa de verosimilitud positiva y para la tasa de verosimulitud negativa, haciendo uso de (1.12) y (1.13), obtenemos las siguientes expresiones:

$$LRP = \frac{\Phi(\frac{\mu_y-c}{\sigma_y})}{\Phi(\frac{\mu_x-c}{\sigma_x})} \qquad LNP = \frac{1-\Phi(\frac{\mu_y-c}{\sigma_y})}{1-\Phi(\frac{\mu_x-c}{\sigma_x})}.$$

Por último, para el valor predicitivo positivo y el valor predictivo negativo definidos en (1.16), tenemos:

$$VPP = \frac{p\Phi\left(\frac{\mu_y-c}{\sigma_y}\right)}{p\Phi\left(\frac{\mu_y-c}{\sigma_y}\right) + (1-p)\left(1-\Phi\left(\frac{\mu_x-c}{\sigma_x}\right)\right)} \qquad VPN = \frac{(1-p)\left(1-\Phi\left(\frac{\mu_x-c}{\sigma_x}\right)\right)}{p\Phi\left(\frac{\mu_y-c}{\sigma_y}\right) + (1-p)\left(1-\Phi\left(\frac{\mu_x-c}{\sigma_x}\right)\right)}.$$

Capítulo 2

Estimación de la curva *ROC*

En este capítulo se hace una revisión de los distintos métodos de estimación de la curva *ROC*, tanto empíricos, paramétricos como no paramétricos. Además, estimamos el *AUC* definido en (1.8) tanto por métodos paramétricos como no paramétricos y para terminar, introducimos unos contrastes básicos para curvas *ROC*. Para el desarrollo de este capítulo, nos hemos basado principalmente en el libro de Krzanowski y Hand, [8].

2.1. Método empírico

El método empírico hace uso de las funciones de distribución empíricas $\hat{F}$ y $\hat{G}$ para la construcción de la curva *ROC*.

Definición 16. *Dada una muestra aleatoria simple $x_1,\dots,x_n$ con función de distribución F, se define la función de distribución empírica asociada a la muestra como:*

$$\begin{array}{rccl} \hat{F}_n\colon & \mathbb{R} & \longrightarrow & [0,1] \\ & c & \longmapsto & \hat{F}_n(c)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(x_i\leq c). \end{array}$$

donde $\mathbb{1}(x_i\leq c)=\begin{cases} 1 \text{ si } x_i\leq c \\ 0 \text{ si } x_i>c. \end{cases}$ *Es decir, para un valor umbral c fijado, mide la proporción de observaciones menores o iguales que c .*

Es bien sabido que, $n\hat{F}_n(c)$ se distribuye como una variable aleatoria Binomial con $n\hat{F}_n(c)\backsim Bin(n,F(c))$.

Así, en nuestro caso tenemos dos muestras aleatorias simples $(x_1,\dots,x_{n_x})$ e $(y_1,\dots,y_{n_y})$ de X e Y tal que:

$$\hat{F}_{n_x}(c)=\frac{1}{n_x}\sum_{i=1}^{n_x}\mathbb{1}(x_i\leq c),\qquad \hat{G}_{n_y}(c)=\frac{1}{n_y}\sum_{i=1}^{n_y}\mathbb{1}(y_i\leq c)\qquad \text{con } n_x+n_y=n. \tag{2.1}$$

Definición 17. *Se define la curva ROC empírica como:*

$$\widehat{ROC}_n(t)=1-\hat{G}_{n_y}(1-\hat{F}_{n_x}^{-1}(1-t)). \tag{2.2}$$

Es también inmediato, que los estimadores empíricos de S y de $1-E$ son:

$$(1-\hat{E})(c)=1-\frac{1}{n_x}\sum_{i=1}^{n_x}\mathbb{1}(x_i\leq c),\qquad \hat{S}(c)=1-\frac{1}{n_y}\sum_{i=1}^{n_y}\mathbb{1}(y_i\leq c). \tag{2.3}$$

La estimación de la curva *ROC* que proporciona el método empírico, es una curva escalonada. Las líneas son horizontales si hay un nuevo falso positivo, un escalón vertical significa que hay un verdadero positivo y un trazo diagonal significa que un nuevo individuo ha pasado a ser falso positivo y otro verdadero positivo, al cambiar el valor de c. Conforme aumenta el tamaño de la muestra los escalones tienden a suavizarse, convergiendo así a la curva *ROC* poblacional.

2.1.1. Intervalos de confianza para el método empírico

Consideremos un punto $(1-\hat{E},\hat{S})$ en la curva ROC empírica, que estima el verdadero punto $(1-E,S)$. Podemos prever tres tipos distintos de variabilidad del punto estimado:

- Variación vertical en $\hat{S}$, sobre muestras que tienen una puntuación fija para $1-E$.
- Variación horizontal en $1-\hat{E}$, sobre muestras que tienen una puntuación fija para S.
- variación bidimensional en $(1-\hat{E},\hat{S})$ sobre muestras que tienen un umbral del punto de corte c fijo.

Trataremos únicamente este último caso, ya que por (2.3) conocemos la expresión de $1-\hat{E}(c)$ y $\hat{S}(c)$ y suponemos que las muestras de las dos poblaciones son independientes entre sí. Así, bajo esta última suposición, $(1-\hat{E},\hat{S})$ son estimaciones independientes basadas en probabilidades binomiales, por lo que los métodos binomiales pueden aplicarse directamente para obtener intervalos de confianza para $1-E$ y para S. Dado que las dos muestras son independientes, la regla de multiplicación de probabilidades establece que si se desea un intervalo de confianza del $100(1-\alpha)\,\%$ para $(1-E,S)$, este viene dado por el rectángulo cuyos lados son los intervalos de confianza del $100(1-\hat{\alpha})\,\%$ para $1-\hat{E}$ y $\hat{S}$ por separado, donde $\hat{\alpha}=1-\sqrt{1-\alpha}$. Sin embargo, si los individuos de las dos muestras no son independientes, no se dispone de un enfoque analítico inmediato.

2.2. Método paramétrico

El método paramétrico asume que la distribución de la variable de decisión o de la variable respuesta de la prueba en cada grupo pertenecen a una familia de distribuciones. Hay que tener en cuenta que la familia de distribuciones no tiene porqué ser la misma en ambos grupos.

Al contrario que en el método empírico, en el caso paramétrico se obtiene una curva suavizada, pero se corre el riesgo de cometer un error importante si la elección de la distribución no es la correcta. Para ello es recomendable realizar previamente un contraste de hipótesis de bondad de ajuste sobre las distribuciones elegidas.

En la sección 1.4, hemos introducido un primer modelo paramétrico basado en distribuciones normales para cada variable. La elección del estimador binormal para ajustar una curva *ROC* suele estar justificada por consideraciones teóricas, entre ellas, la familiaridad con el modelo normal o porque ofrece buenos resultados aún teniendo una ciertas observaciones que no siguen una distribución normal.

Además, debemos intentar aplicar transformaciones monótonas (por ejemplo de tipo Box-Cox) si nuestros datos no se ajustan a la distribución deseada, ya que como hemos visto en la Proposición 1.3, la curva *ROC* es invariante bajo este tipo de transformaciones. Existen varios métodos que ajustan la variable a familias de distribuciones como la logística o la exponencial negativa, pero la distribución más usada es la normal.

Por tanto, si nos centramos en el enfoque binormal, tal y como hemos definido la curva *ROC* en (1.17), debemos estimar los parámetros a y b y la forma de estimarlos es a través del estimador máximo verosimil.

Un enfoque analítico, como se muestra en [8], se concibió originalmente para su uso en datos categóricos ordenados, en los que el estadístico de clasificación (defindo como R) puede tomar sólo uno de un conjunto finito de valores o categorías $C_1,C_2,\ldots,C_k$. Dicho enfoque asume que existe una variable W y un conjunto de valores umbrales desconocidos $-\infty=w_0<w_1<\ldots<w_k=\infty$ tal que R pertenece a la categoria C_i si y solo si $w_{i-1}<W\leq w_i$. Si identificamos $W=h(R)$ como el resultado de una transformación desconocida h de R a la normalidad, entonces se deduce que:

$$p_{ix}=P(R\in C_i|Z=0)=\Phi(\frac{w_i-\mu_x}{\sigma_x})-\Phi(\frac{w_{i-1}-\mu_x}{\sigma_x})$$

$$p_{iy}=P(R\in C_i|Z=1)=\Phi(\frac{w_i-\mu_y}{\sigma_y})-\Phi(\frac{w_{i-1}-\mu_y}{\sigma_y}).$$

Si n_{ix}, n_{iy} son el número de observaciones de individuos de las ambas poblaciones, respectivamente, que caen en la categoría C_i podemos escribir el logaritmo de la función de verosimilitud de tipo multinomial como:

$$log\mathscr{L} = \sum_{i=1}^{k} (n_{ix} log(p_{ix}) + n_{iy} log(p_{iy})). \tag{2.4}$$

Estos mismos autores desarrollaron el algoritmo iterativo para maximizar (2.4) con respecto a los parámetros del modelo en presencia de datos categóricos, y sigue siendo uno de los método más populares.

2.2.1. Intervalos de confianza en métodos paramétricos

Si se ha asumido una distribución específica para cada una de las dos poblaciones y los parámetros del modelo se han estimado por máxima verosimilitud, entonces la distribución conjunta asintótica de los parámetros se obtiene a partir de la teoría de máxima verosimilitud. En concreto, si $\theta = (\theta_1, ..., \theta_p)^T$ es el vector de parámetros desconocidos del modelo y $log\mathscr{L}$ es la log-verosimilitud de la muestra, entonces el estimador de máxima verosimilitud $\hat{\theta}$ se distribuye asintóticamente como una vector normal multivariante con media θ y matriz de dispersión I^{-1} en la que el elemento (i, j) de I viene dado por $-E(\frac{\partial^2 log\mathscr{L}}{\partial\theta_i \partial\theta_j})$. La estimación paramétrica de la curva *ROC* es una función de todos los parámetros estimados del modelo, por lo que podemos utilizar el resultado anterior, el hecho de que las dos muestras son independientes y el método delta para obtener la varianza y, por tanto, el error estándar de cualquier punto de la curva *ROC*. Los parámetros desconocidos se sustituyen de nuevo por sus estimaciones de máxima verosimilitud y, así, se llega a poder obtener intervalos de confianza punto a punto.

El modelo binormal permite una expresión relativamente sencilla para un intervalo de confianza al $100(1-\alpha)\,\%$. Para este modelo, el valor $\hat{S}$ viene dado por $\hat{S} = \Phi(\hat{a} + \hat{b} z_x)$ donde $\hat{a}$ y $\hat{b}$ son las estimaciones de máxima verosimilitud de a y b. La varianza de $\hat{a} + \hat{b} z_x$ viene dada por

$$V = Var(\hat{a}) + z_x^2 Var(\hat{b}) + 2 z_x cov(\hat{a}, \hat{b}) \tag{2.5}$$

y esto puede obtenerse fácilmente porque las varianzas de $\hat{a}$ y $\hat{b}$, y la covarianza entre ellas son los elementos de la inversa de la matriz hessiana evaluada en $\hat{a}, \hat{b}$. Los límites de confianza del $100(1-\alpha)\,\%$ para $a + b z_x$ están entonces dados por

$$\hat{a} + \hat{b} z_x \pm \hat{V}^{\frac{1}{2}} z_{1-\frac{\alpha}{2}} \tag{2.6}$$

y esto da límites de confianza puntuales en valores especificos de $1 - E$ para $S = \Phi(a + b z_x)$, con $z_x = \Phi^{-1}(1-E) = \frac{\mu_x - c}{\sigma_x}$.

2.3. Cálculo del *AUC*

Basándonos en la expresión dada en (1.8) del área bajo la curva, la estimación de esta cantidad a partir de los datos de la muestra es inmediata. Se puede estimar la curva *ROC* de forma empírica o ajustando una curva suave utilizando el método paramétrico, y luego obtener $\widehat{AUC}$ por integración numérica. Esto último puede llevarse a cabo mediante la regla del trapecio compuesta (véase [3, Capítulo 7]), aunque esto produce una pequeña subestimación de las curvas *ROC*.

En el caso particular de que la curva *ROC* se haya estimado empíricamente, entonces la integración numérica es innecesaria ya que sabemos por (1.9) que $AUC = P(Y > X)$. Así, esto nos conduce a la equivalencia señalada por Bamber en [1] de que el área bajo la curva *ROC* empírica es igual al estadístico U de Mann-Whitney. Una definición formal de este estadístico es

$$U = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} [\mathbb{1}(y_j > x_i) + \frac{1}{2} \mathbb{1}(y_j = x_i)] \tag{2.7}$$

que se puede ver como la proporción de pares en los que la puntuación del individuo de la muestra que no presenta una cierta característica supera a la del individuo de la muestra que sí que presenta la característica, y la mitad de la proporción de empates. Dado que la estadística de clasificación es continua, siempre que la escala de medición sea lo suficientemente precisa, podemos suponer una probabilidad insignificante de obtener empates, en cuyo caso se deduce que $E(U) = P(Y > X) = AUC$, y que U proporciona un estimador insesgado de AUC.

En el caso de los modelos paramétricos basados en supuestos de distribuciones específicas, el AUC suele poder obtenerse analíticamente en términos de los parámetros de la población, por lo que se estima simplemente sustituyendo las estimaciones de máxima verosimilitud de estos parámetros. Aunque también, siempre se puede recurrir a la integración numérica para estimar el AUC.

2.3.1. Intervalos de confianza para el *AUC*

Para obtener intervalos de confianza para el AUC, necesitamos tener una estimación de la varianza de $\widehat{AUC}$.

En el caso de los métodos paramétricos, la teoría de la máxima verosimilitud proporcionará expresiones asintóticas para las varianzas y covarianzas de los parámetros y el método delta proporcionará la varianza necesaria (como en el caso de los intervalos de confianza para la curva *ROC*).

Por lo tanto, nos centramos aquí en los enfoques no paramétricos. Consideremos en primer lugar la curva *ROC* empírica y el estadístico U Mann-Whitney dado en (2.7). La expresión asintótica para su varianza es la dada por Hanley y McNeil en [6]:

$$Var(\widehat{AUC}) = \frac{1}{n_x n_y}(\widehat{AUC}(1-\widehat{AUC}) + (n_y-1)(Q_1 - \widehat{AUC}^2) + (n_x-1)(Q_2 - \widehat{AUC}^2)) \tag{2.8}$$

donde Q_1 es la probabilidad de que las puntuaciones de clasificación de dos individuos elegidos al azar de la población que presentan una cierta característica superen la puntuación de un individuo elegido al azar de la población que no la presenta, y Q_2 es la probabilidad inversa de que la puntuación de clasificación de un individuo elegido al azar de la población que la presenta supere ambas puntuaciones de dos individuos elegidos al azar de la población que no presenta la característica.

Si se estiman los parámetros $\widehat{AUC}$, $\hat{Q}_1$ y $\hat{Q}_2$, se obtiene una estimación de $Var(AUC)$. La estimación de $\hat{Q}_1$ y $\hat{Q}_2$ es sencilla, por ejemplo, $\hat{Q}_1$ viene dado por la consideración de todas las posibles tripletas que comprenden dos miembros que presenten la característica y uno que no la presente, y encontrando la proporción en la que las dos puntuaciones de los que la presentan superan la puntuación del que no la presenta. Por lo que el intervalo de confianza del $100(1-\alpha)\%$ para el AUC viene dado por $\widehat{AUC} \pm \sqrt{\widehat{Var}(\widehat{AUC})}z_{1-\frac{\alpha}{2}}$.

2.4. Método de regresión binaria

Se ha estudiado recientemente una forma menos directa de estimar $F(x)$ y $G(x)$, buscando primero las estimaciones de $P(Z=1|T=c)$ y $P(Z=0|T=c)$. Así, definimos $\pi_j(c) = P(Z=j|T=c)$ para $j=0,1$. Además, se denota la probabilidad global de que un individuo de la muestra provenga de $Z=j$ por π_j para $j=0,1$. En [9] se propone reescribir los dos conjuntos de valores de la muestra en la forma simple como pares (u_i, c_i) para $i = 1,2,...,n$, donde $u_i = 0,1$ según si la i-ésima observación proceda de la población con una cierta característica o no, y c_i es el valor del estadístico de clasificación para el i-ésimo individuo.

Así, la verosililitud conjunta $\mathscr{L}$ de los pares se puede calcular:

$$\begin{aligned} \mathscr{L} &= \mathscr{L}[(u_1,c_1),\ldots,(u_n,c_n)] = \mathscr{L}[u_1,\ldots,u_n|c_1,\ldots,c_n]\mathscr{L}[c_1,\ldots,c_n] = \\ &= (\prod_{i=1}^{n}\pi_0(c_i)^{1-u_i}\pi_1(c_i)^{u_i})(\prod_{i=1}^{n}[\pi_0 f(c_i) + \pi_1 g(c_i)]). \end{aligned}$$

Se proponen utilizar métodos de regresión logística para estimar el $\pi_j(c)$ en términos de cualquier conjunto de funciones base. Por ejemplo, en [8] se cita un trabajo donde los autores utilizaron un modelo logístico para el $\pi_j(c)$, con

$$\pi_j(c) = \frac{e^{\alpha^* + \beta^T r(c)}}{1 + e^{\alpha^* + (\beta)^T r(c)}}, \tag{2.9}$$

donde $r(c)$ es un vector de funciones base. Habiendo mostrado en un artículo anterior que

$$\frac{g(c)}{f(c)} = e^{[\alpha + \beta^T r(c)]}$$

donde $\alpha = \alpha^* + log(\frac{\pi_0}{\pi_1})$ y esto proporciona una conexión entre las dos funciones de densidad.

Estos artículos recomiendan el uso de métodos bootstrap para calcular intervalos de confianza tanto para la curva *ROC* como para los principales índices, p. ej. el *AUC*.

2.5. Método no paramétrico suavizado

El método no paramétrico para la construcción de la curva *ROC* no hace suposición alguna sobre la distribución de los resultados de la prueba en ambos grupos. Los enfoques descritos en los modelos paramétricos comparten algunas debilidades, principalmente que la distribución asumidas puede ser inapropiada para los datos en cuestión. En consecuencia, varios autores han recurrido más recientemente a métodos no paramétricos, que son aplicables de forma muy general. Tener en cuenta que, haciendo uso de la función kernel para estimar las distribuciones de la variable de decisión en ambos grupos se obtiene una curva suavizada.

El primer enfoque de este tipo fue el dado en [16], donde se sugirió utilizar métodos de densidad kernel (ver [14] como referencia básica de estos métodos) para estimar las funciones de densidad en cada población, seguido de la integración de las funciones de densidad para obtener estimaciones de $F(x)$ y $G(x)$.

Sean $(x_1, ..., x_{n_x})$ e $(y_1, ..., y_{n_y})$ dos muestras independientes de X e Y , respectivamente. Los estimadores kernel de f y g son

$$\hat{f}(c) = \frac{1}{n_x h_x} \sum_{i=1}^{n_x} K_1\left(\frac{c - x_i}{h_x}\right), \qquad \hat{g}(c) = \frac{1}{n_y h_y} \sum_{i=1}^{n_y} K_2\left(\frac{c - y_i}{h_y}\right). \tag{2.10}$$

La elección de una de las muchas funciones kernel disponibles es relativamente poco importante, ya que todas dan resultados comparables como se cita en [5]. Además, las funciones kernel cumplen unas ciertas propiedades:

$$\int_{\mathbb{R}} K_i(c)dc = 1, \qquad \int_{\mathbb{R}} c \cdot K_i(c)dc = 0, \qquad \int_{\mathbb{R}} c^2 \cdot K_i(c)dc > 0 \tag{2.11}$$

para $i = 1, 2$.

Por otra parte, tener en cuenta que h_x y h_y es una secuencia de números positivos llamados anchos de banda o bandwidths que determina cuánto de suavizada queda la curva y hay que tener especial cuidado en elegir cual es el valor de h_x y h_y.

Utilizando estos estimadores, las funciones de distribución pueden estimarse como

$$\hat{F}(c) = \frac{1}{n_x h_x} \sum_{i=1}^{n_x} \int_{-\infty}^{x} K_1\left(\frac{c - x_i}{h_x}\right), \qquad \hat{G}(c) = \frac{1}{n_y h_y} \sum_{i=1}^{n_y} \int_{-\infty}^{x} K_2\left(\frac{c - y_i}{h_y}\right). \tag{2.12}$$

Estas integrales se pueden calcular de manera numérica. y así se puede estimar directamente la curva *ROC*.

Una opción popular para usar como kernel es el kernel Gaussiano, y en este caso podemos escribir las funciones de distribución defindas en (2.12) como

$$\hat{F}(c) = \frac{1}{n_x h_x} \sum_{i=1}^{n_x} \Phi\left(\frac{c - x_i}{h_x}\right) \qquad \hat{G}(c) = \frac{1}{n_y h_y} \sum_{i=1}^{n_y} \Phi\left(\frac{c - y_i}{h_y}\right). \tag{2.13}$$

En [16] se sugirió utlizar la siguiente función kernel:

$$K_i\left(\frac{c-a}{b}\right) = \begin{cases} \frac{15}{16}[1 - (\frac{c-a}{b})^2]^2 & \text{, si } c \in (a-b, a+b) \\ 0 & \text{, si } c \notin (a-b, a+b) \end{cases}. \tag{2.14}$$

Se utilizó dicha función con el objetivo de definir los anchos de banda (ver [14]) como $h_x = 0.9\,\text{mín}\frac{(sd_x, iqr_x)}{n_x^{\frac{1}{5}}}$ y $h_y = 0.9\,\text{mín}\frac{(sd_y, iqr_y)}{n_y^{\frac{1}{5}}}$. Donde sd, iqr indican la desviación típica de la muestra y el rango intercuartilico para las respectivas poblaciones.

Sin embargo, su elección de ancho de banda no es óptima para la curva *ROC*, ya que ésta depende de las funciones de distribución $F(x), G(x)$ y la optimización para estimar las funciones de densidad no implica la optimización para estimar las funciones de distribución. Por ello, se investigaron varios anchos de banda diferentes y este estudio mejoró la elección anterior, al obtener estimaciones asintóticamente óptimas de las funciones de distribución $F(x)$ y $G(x)$.

Generalmente se necesita la integración numérica para estimar el *AUC*. En [8], se indica que la estimación kernel resultante del *AUC* puede expresarse como

$$\widehat{AUC} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{i=1}^{n_y} \Phi\left(\frac{y_j - x_i}{\sqrt{h_x^2 + h_y^2}}\right). \tag{2.15}$$

2.5.1. Intervalos de confianza

En [16] obtienen intervalos de confianza para S dado un valor fijo de $1-E$, y para $(1-E, S)$ para un valor fijo c del umbral del estadístico de clasificación, siguiendo su estimación no paramétrica de la curva *ROC* basada en el método kernel. Para obtener los intervalos, primero utilizan transformaciones *logit* de todas las probabilidades, basándose en que una aproximación normal funciona mejor en los valores transformados. Así, utilizan intervalos de confianza basados en la normalidad para los valores transformados y, a continuación, utilizan la transformación inversa en los límites de los intervalos para obtener intervalos de confianza para los valores originales. La transformada logit de la probabilidad θ viene dada por

$$\phi = logit(\theta) = log\left(\frac{\theta}{1-\theta}\right), \tag{2.16}$$

de modo que la transformada inversa es $\theta = logit^{-1}(\phi) = \frac{1}{1+e^{-\phi}}$. Tomamos $u = logit(1-E)$, $v = logit(S)$, $\hat{u} = logit(1-\hat{E})$, y $\hat{v} = logit(\hat{S})$.

Nos centramos en un intervalo de confianza para $(1-E, S)$ en un valor fijo del umbral del estadístico de clasificación. Tenemos que $\hat{u}$, $\hat{v}$ son asintóticamente normales de forma independiente con medias u, v y varianzas estimadas por $\widehat{Var}(\hat{u}) = \frac{1}{n_x(1-\hat{E})\hat{E}}$ y $\widehat{Var}(\hat{v}) = \frac{1}{n_y(1-\hat{S})\hat{S}}$. Los intervalos de confianza $100(1-\alpha)\%$ para cada una de las u y v vienen dados por $\hat{u} \pm \sqrt{\widehat{Var}(\hat{u})} z_{1-\frac{\alpha}{2}}$ y $\hat{v} \pm \sqrt{\widehat{Var}(\hat{v})} z_{1-\frac{\alpha}{2}}$, y como $\hat{u}$, $\hat{v}$ son independientes, entonces estos intervalos definen una región de confianza rectangular para (u, v) que tiene un coeficiente de confianza $100(1-\alpha)^2$.

Con respecto a las estimaciones del *AUC*, en [16] se comenta que el suavizado introducido por la estimación kernel no afecta al verdadero error estándar en un primer orden de aproximación, por lo que los valores de $Var(\widehat{AUC})$ dados en (2.8) pueden utilizarse también para el área bajo la curva suavizada. Sin embargo, sugieren que cualquier intervalo de confianza debe calcularse para el *AUC* transformado y luego retrotransformado a la escala original, con el fin de garantizar que se mantenga dentro del rango $(0, 1)$.

2.6. Contrastes aplicados a las curvas *ROC*

Ya hemos visto varias formas de estimar tanto la curva *ROC* como uno de los índices más importantes como es el *AUC*. Una vez estimadas las curvas, nos planteamos a través de un estadistico de contraste si podemos aceptar o rechazar valores específicos del *AUC* frente a sus estimaciones.

2.6.1. Contraste para una prueba

Una vez obtenido el valor del *AUC* para una prueba es recomendable contrastar la posibilidad de que la variable de decisión sea aleatoria y no haya salido reflejado en el cálculo del *AUC* debido a la variabilidad de la muestra. Este constraste viene dado por:

$$\begin{cases} H_0 : AUC = 0.5 \\ H_1 : AUC \neq 0.5 \end{cases}$$

El estadístico que resuelve tal contraste fue expuesto por Hanley y McNeil en [7] y es:

$$z = \frac{\widehat{AUC} - 0.5}{\sqrt{\widehat{Var}(\widehat{AUC})}}.$$

Dicho estadístico sigue una distribución $N(0,1)$ bajo la hipótesis nula. Así, para un nivel de significación α se rechaza la aleatoriedad de nuestra prueba si $\mid z \mid > z_{\alpha/2}$, donde $z_{\alpha/2}$ es el cuantil de la distribución $N(0,1)$. Es recomendable que los resultados obtenidos vengan acompañados con una gráfica de la curva.

2.6.2. Contraste para dos pruebas

Dadas dos pruebas diagnósticas, se considera mejor en el sentido discriminante, la que mayor área tenga y si ambas áreas fuesen iguales, podría ser porque son la misma curva *ROC* o porque una nos ofrece una prueba más sensible y otra más específica.

Así, se nos presenta el problema de si una prueba tiene mayor área que otra única y exclusivamente por tener mayor capacidad discriminante o es debido a la variabilidad de la muestra. Por tanto, planteamos el siguiente contraste:

$$\begin{cases} H_0 : AUC_1 = AUC_2 \\ H_1 : AUC_1 \neq AUC_2 \end{cases}$$

En [7], se propone el siguiente estadístico para resolver este contraste:

$$z = \frac{\widehat{AUC_1} - \widehat{AUC_2}}{\sqrt{\widehat{Var}(\widehat{AUC_1} - \widehat{AUC_2})}},$$

donde $\sqrt{\widehat{Var}(\widehat{AUC_1} - \widehat{AUC_2})} = \sqrt{\widehat{Var}(\widehat{AUC_1}) + \widehat{Var}(\widehat{AUC_2}) - 2\widehat{Cov}(\widehat{AUC_1}, \widehat{AUC_2})}$.

Dicho estadístico también sigue una distribución $N(0,1)$ bajo la hipótesis nula. Hay que tener en cuenta que este contraste es para rechazar o no la igualdad de áreas y dos curvas *ROC* muy diferentes podrían tener igual área. Así, en caso de aceptar la hipótesis nula, sería recomendable acompañar el resultado con un examen visual de las curvas.

Capítulo 3

Extensiones a covariables y aplicación práctica

En la primera parte del capítulo se introduce la curvas *ROC* condicionadas a covariables y se consideran diversas formas de estimación. Para esta sección, nuestro articulo fuente ha sido [10]. En la segunda parte del capítulo, se plantea una aplicación a datos reales donde el objetivo es mostrar con mayor claridad de una forma visual los conceptos previamente descritos. En particular, nos centramos en detalle en el estudio de curvas *ROC* dependientes de covariables discretas y continuas.

3.1. Curva *ROC* y covariables

A menudo resulta que una nueva variable o conjunto de variables proporciona información clasificatoria útil que modifica el comportamiento del clasificador en una o ambas poblaciones. Estas variables adicionales se denominan covariables, y se deben incorporar a cualquier análisis en el que intervenga el clasificador.

En muchas situaciones el rendimiento de una prueba diagnóstica y, por tanto, su capacidad discriminatoria pueden verse afectados por las covariables. En [10, p.23], se ofrecen varios ejemplos de covariables que pueden afectar al resultado de una prueba. Por ejemplo, en epimediología, las características del paciente, como la edad y el sexo, son covariables importantes que hay que tener en cuenta. La incorporación de covariables en la curva *ROC* puede hacerse con dos fines:

- Para obtener curvas *ROC* específicas de las covariables, o curvas *ROC* que condicionan un valor específico de un vector de covariables (sección 3.1.1) . A su vez, podemos seguir dos enfoques distintos:
  - El ajuste directo en el que el efecto de las covariables se modela en la propia curva *ROC*.
  - El ajuste indirecto, en el que el efecto de las covariables en las distribuciones se modela primero en las dos poblaciones y la curva *ROC* se calcula después de las distribuciones modificadas.
- Para obtener algún tipo de curva *ROC* ajustada por covariables, que tiene en cuenta la información de las covariables de cada punto de datos para obtener una mejor medida de la capacidad discriminatoria que la curva *ROC* marginal (véase sección 3.1.2).

Supongamos que, junto con las variables continuas X e Y, también se dispone de vectores de covariables M. Aunque, en algunas ocasiones, podría ser interesante evaluar la capacidad discriminatoria de una prueba diagnóstica con respecto a covariables específicas para cada población, asumiremos covariables comunes a ambas poblaciones.

Definición 18. *La curva ROC condicional o específica de una covariable, con valor m, se define como*

$$ROC_m(t) = 1 - G(F^{-1}(1-t|m)|m) \tag{3.1}$$

con

$$F(c|m) = P(X \leq c|m) \qquad G(c|m) = P(Y \leq c|m). \tag{3.2}$$

Obsérvese que en este caso se pueden obtener varias curvas *ROC* diferentes para cada valor m en la covariable común M.

Asociadas a la curva *ROC* condicional, también pueden definirse algún índice como es el área bajo la curva.

Definición 19. *Se define el AUC condicional como*

$$AUC_m = \int_0^1 ROC_m(t)dt \tag{3.3}$$

de forma equivalente a lo visto en (1.9).

3.1.1. Estimación de la curva *ROC* condicional

Supongamos que tenemos dos muestras independientes de observaciones idénticamente distribuidas $(x_1, m_{x,1}), \ldots, (x_{n_x}, m_{x,n_x})$ de la población (X,M) y $(y_1, m_{y,1}), \ldots, (y_{n_y}, m_{y,n_y})$ de la población (Y,M) con $n_x + n_y = n$.

Estimadores basados en funciones de distribución condicional.

Un estimador de la curva *ROC* condicional se obtiene directamente de (3.1). Dado el valor m de la covariable, el estimador puede construirse como:

$$\widehat{ROC_m}(t) = 1 - \hat{G}(\hat{F}^{-1}(1-t|m)|m) \tag{3.4}$$

donde $\hat{F}(t|m)$ y $\hat{G}(t|m)$ son estimadores de las distribuciones condicionales $F(t|m)$ y $G(t|m)$, respectivamente. Cuando restringimos nuestra atención a las covariables unidimensionales, las distribuciones condicionales pueden estimarse de forma no paramétrica, por ejemplo, mediante los estimadores basados en el método kernel dados en [10, p.29]:

$$\hat{F}_{h_x}(c|m) = \frac{\sum_{i=1}^{n_x} k(\frac{m-m_{x,i}}{h_x}) \mathbb{1}(x_i \leq c)}{\sum_{i=1}^{n_x} k(\frac{m-m_{x,i}}{h_x})}, \quad \hat{G}_{h_y}(c|m) = \frac{\sum_{i=1}^{n_y} k(\frac{m-m_{y,i}}{h_y}) \mathbb{1}(y_i \leq c)}{\sum_{i=1}^{n_y} k(\frac{m-m_{y,i}}{h_y})} \tag{3.5}$$

donde k es una función kernel con h_x y h_y los parámetros de suavización. Bajo este enfoque, el estimador de la curva *ROC* condicional para un valor de específico de la covariable utiliza la información correspondiente de los individuos cuyos valores de la covariable están cerca de m.

El estimador dado en (3.4) al ser empírco hereda las discontinuidades de $\hat{G}$ y $\hat{F}$. En [10, p.29], se cita un trabajo donde se obtiene un estimador suave no paramétrico de la curva *ROC* condicional. Se propone suavizar la curva *ROC* empírica mediante técnicas kernel y la versión suavizada de la curva condicional propuesta es:

$$\widehat{ROC}_{m,h}(t) = 1 - \int_{\mathbb{R}} \hat{G}_{h_y}(\hat{F}_{h_x}^{-1}(1-t+hu|m)|m)k(u)du \tag{3.6}$$

donde el parámetro h controla la cantidad de suavizado. En dicho trabajo se propone un método bootstrap para elegir los parámetros de suavizado.

Estimadores basados en la metodología de regresión directa

Una forma alternativa de incorporar la información de las covariables al análisis *ROC* es mediante modelos de regresión. En la metodología directa el efecto de las covariables se evalúa directamente en la curva *ROC*. Así, expresemos la curva *ROC* condicional dada en (3.1) como:

$$\begin{aligned} ROC_m(t) &= 1-G(F^{-1}(1-t|m)|m) = 1-P(Y \leq F^{-1}(1-t|m)|M=m) = \\ &= 1-P(F(Y|m) \leq 1-t|M=m) = P(1-F(Y|m) < t|M=m) = \\ &= E(\mathbb{1}(1-F(Y|m) < t)|M=m). \end{aligned} \tag{3.7}$$

Por tanto, la curva *ROC* condicional puede verse como la función de distribución condicional de la variable aleatoria $1-F(Y|m)$ o como el valor esperado condicional de la variable binaria $\mathbb{1}(1-F(Y|m) < t)$.

Estas dos interpretaciones justifican que se exprese la curva *ROC* condicional como una especie de modelo de regresión de la forma

$$ROC_m(t) = \gamma(\mu(m), \Gamma(t)), \tag{3.8}$$

donde γ es una función bivariante en $[0,1]$ y Γ es una función definida en el intervalo $[0,1]$, que está relacionada con la forma de la curva *ROC*. La función μ recoge el efecto de las covariables sobre la curva *ROC* condicional. Para obtener un modelo válido de curvas *ROC*, es necesario imponer algunas restricciones. En particular, la función γ debe ser monótona y creciente en t, con $\gamma(\mu(m), \Gamma(0)) = 0$ y $\gamma(\mu(m), \Gamma(1)) = 1$ para todo m.

El modelo dado en (3.8) representa la fórmula general de la metodología directa aunque solo la especificación aditiva ha sido tratada en la literatura estadística, es decir,

$$ROC_m(t) = \gamma(\mu(m) + \Gamma(t)). \tag{3.9}$$

En [10, *p*.33], se citan varios articulos donde se difiere en las suposiciones hechas sobre las funciones γ, μ y Γ. Varios de estos artículos, suponen que γ es conocida, que el efecto de las covariables sobre la curva *ROC* condicional es lineal y que la función Γ tiene una forma paramétrica. En otro trabajo se relajan las suposiciones de γ, permitiendo una función γ completamente desconocida.

Estimadores basados en la metodología de regresión inducida

La metodología inducida en el análisis *ROC* consiste en modelar el efecto de las covariables a través de modelos de regresión que vinculan la variable de clasificación y las covariables en cada población por separado. Así, los modelos de regresión se utilizarán entonces para componer la curva *ROC* condicional donde la relación entre la covariable y la variable de clasificación en cada población viene dada por los modelos de regresión con escala y localización:

$$X = \mu_x(M) + \sigma_x(M)\varepsilon_x, \qquad Y = \mu_y(M) + \sigma_y(M)\varepsilon_y \tag{3.10}$$

donde, para $j \in \{x,y\}$, $\mu_j(m)$ y $\sigma_j^2(m)$ son la media condicional y la varianza condicional de X e Y dado $M=m$, respectivamente, y el error ε_j es independiente de la covariable M. La independencia entre el error y la covariable en el modelo de regresión nos permite reescribir las expresiones dadas en (3.2) en términos de la distribución del error de regresión como sigue:

$$\begin{aligned} F(c|m) &= P(X \leq c|M=m) = P(\mu_x(M) + \sigma_x(M)\varepsilon_x \leq c|M=m) = \\ &= P\left(\varepsilon_x \leq \frac{c-\mu_x(m)}{\sigma_x(m)}\right) = H_x\left(\frac{c-\mu_x(m)}{\sigma_x(m)}\right) \end{aligned} \tag{3.11}$$

y análogamente

$$G(c|m) = P(Y \leq c|M=m) = H_y\left(\frac{c-\mu_y(m)}{\sigma_y(m)}\right) \tag{3.12}$$

donde, para $j \in \{x,y\}$, $H_j(c) = P(\varepsilon_j \leq c)$ es la función de distribución del error de regresión. Se puede establecer una relación análoga entre la función cuantil condicional $F^{-1}(-|m)$, $G^{-1}(-|m)$, y la función cuantil de H_j^{-1}, mediante las expresiones $F^{-1}(t|m) = \mu_x(m) + \sigma_x(m)H_x^{-1}(t)$ y $G^{-1}(t|m) = \mu_y(m) + \sigma_y(m)H_y^{-1}(t)$. Por lo tanto, para un valor fijo de la covariable m, y para $0 < t < 1$, la curva *ROC* condicional dada en (3.1), puede expresarse como

$$\begin{aligned} ROC_m(t) &= 1 - G(F^{-1}(1-t|m)|m) = 1 - G(\mu_x(m) + \sigma_x(m)H_x^{-1}(1-t)|m) = \\ &= 1 - H_y\left(\frac{\mu_x(m) + \sigma_x(m)H_x^{-1}(1-t) - \mu_y(m)}{\sigma_y(m)}\right) = 1 - H_y(H_x^{-1}(1-t)b(m) - a(m)) \end{aligned}$$

donde $a(m) = \frac{\mu_y(m)-\mu_x(m)}{\sigma_y(m)}$ y $b(m) = \frac{\sigma_x(m)}{\sigma_y(m)}$. Esta fórmula nos permite expresar la curva *ROC* condicional en términos de la función de distribución y de la función de cuantiles de los errores de regresión, que no son condicionales, por lo que, desde el punto de vista de la estimación, sólo hay que estimar la distribución del error en cada población.

Varios autores, como se muestra en [10, $p.31$], se centraron en un marco no paramétrico, aunque solo consideraban covariables unidimensionales. Sugieren realizar los siguientes pasos cuando los modelos (3.10) se estiman bajo un marco no paramétrico:

1. Para $j \in \{x,y\}$, estimar de forma no paramétrica $\hat{\mu}_j(m)$ y $\hat{\sigma}_j(m)$ con funciones kernel.
2. Estimar de forma empírica la distribución de los errores. Así, tenemos que $\hat{H}_j(c) = \frac{\sum_{i=1}^{n_j} \mathbb{1}(\hat{\varepsilon}_{ji} \leq c)}{n_j}$ donde $\hat{\varepsilon}_{xi} = \frac{x_i - \hat{\mu}_x(m_{x,i})}{\hat{\sigma}_x(m_{x,i})}$ y $\hat{\varepsilon}_{yi} = \frac{y_i - \hat{\mu}_y(m_{y,i})}{\hat{\sigma}_y(m_{y,i})}$.

Por tanto, un estimador empírico de la curva *ROC* condicional es:

$$\widehat{ROC}_m(t) = 1 - \hat{H}_y(\hat{H}_x^{-1}(1-t)\hat{b}(m) - \hat{a}(m)) \tag{3.13}$$

donde $\hat{a}(m) = \frac{\hat{\mu}_y(m)-\hat{\mu}_x(m)}{\hat{\sigma}_y(m)}$ y $\hat{b}(m) = \frac{\hat{\sigma}_x(m)}{\hat{\sigma}_y(m)}$. Como el estimador anterior de la curva *ROC* condicional no es continua, en [10, $pp.31-32$], se propone una versión suavizada:

$$\widehat{ROC}_{m,h}(t) = 1 - \int_{\mathbb{R}} \hat{H}_y(\hat{H}_x(1-t+hu)\hat{b}(m) - \hat{a}(m))k(u)du. \tag{3.14}$$

Los autores muestran que el estimador también admite la siguiente expresión explícita:

$$\widehat{ROC}_{m,h}(t) = \frac{1}{n_y}\sum_{i=1}^{n_y} K\left(\frac{\hat{H}_y(\frac{\hat{\varepsilon}_{yi} - \hat{a}(m)}{\hat{b}(m)}) - 1 + t}{h}\right) \tag{3.15}$$

donde K es la función de distribución correspondiente a la densidad kernel k.

En estos trabajos se detallan las propiedades asintóticas de los estimadores dados en (3.13) y (3.14) y proponen un contraste basado en el método bootstrap para comprobar el efecto de las covariables sobre la curva *ROC* condicional. También presentan un estimador condicional para el *AUC* siendo la integral aproximada por métodos numéricos.

3.1.2. Estimación de la curva *ROC* ajustada por covariables

La curva *ROC* condicional representa la capacidad discriminatoria de una prueba pero para valores específicos del vector de covariables. Sin embargo, sería de gran interés disponer de medidas discriminatorias globales que también tengan en cuenta la información de las covariables.

Definición 20. *La curva ROC ajustada a las covariables se define como un promedio de curvas ROC condicionales ponderadas según la distribución de la covariable en la población con una cierta característica, es decir:*

$$AROC(t) = \int_{\mathbb{R}} ROC_m(t)d\Omega_y(m) \tag{3.16}$$

donde $\Omega_y(m) = P(M \leq m|Z = 1)$ *es la función de distribución multivariante del vector* $M|Z = 1$.

En [10, *p*.34], se cita un trabajo en el que se admiten otras representaciones equivalentes a (3.16).

$$AROC(t) = P(Y > F^{-1}(1-t|(M|Z=1)) \tag{3.17}$$

y proponen estimadores de la forma

$$\widehat{AROC}(t) = \frac{1}{n_y}\sum_{i=1}^{n_y} \mathbb{1}(y_i > \hat{F}^{-1}(1-t|m_{y,i})) \tag{3.18}$$

donde $\widehat{F}^{-1}(1-t|m_{y,i})$ puede estimarse de forma no paramétrica. En el contexto de la metodología inducida descrita en (3.1.1), como se menciona en [10, *p*.35], se puede hacer uso de la relación entre el cuantil condicional y el cuantil de los errores de regresión para obtener el siguiente estimador no paramétrico

$$\widehat{AROC}(t) = \frac{1}{n_y}\sum_{i=1}^{n_y} \mathbb{1}\left(\frac{y_i - \hat{\mu}_x(m_{y,i})}{\hat{\sigma}_x(m_{y,i})} > \hat{H}_x^{-1}(1-t)\right) \tag{3.19}$$

con la notación de los estimadores introducidos anteriormente.

3.2. Aplicación a un caso real

Por último nos centraremos en realizar un caso práctico con datos aplicando todo lo visto hasta ahora pero especialmente haciendo un estudio de la curva *ROC* dependiente de covariables. Nuestro conjunto de datos, el cual se llama `endosim`, se encuentra en el paquete `npROCRegression` [13] y se utilizaron con el objetivo de relacionar el Índice de Masa Corporal (*IMC*) para detectar pacientes con mayor riesgo de problemas cardiovasculares (la variable toma valor 1 si se tienen más de 2 factores de riesgo y 0 si no se tienen), comprobando el posible efecto de la edad y el género.

En nuestro conjunto de datos tenemos:

- 2840 individuos donde 1317 son hombres y 1523 son mujeres.
- 2149 individuos no tienen riesgo y 691 sí que tienen riesgo.
- La media del *IMC* de las personas que tiene riesgo es 30.15 frente a 25.66 de los que no tienen riesgo.
- La media de edad de los individuos que no tienen riesgo es de 38.18 y de 51.54 de los que sí que lo tienen.

Cabe destacar que no podemos eliminar ningún dato atípico ya que los posibles valores del *IMC* no parecen estar muy lejos de los valores más comunes.

Nos vamos a plantear primero, una curva *ROC* donde nuestra variable Z representa si el individuo tiene más de 2 factores de riesgo cardiovascualares y la variable T es el *IMC*. Hay que tener en cuenta que nos podriamos plantear una curva *ROC* en la que nuestra variable T sea la edad. El estudio sería semejante al que vamos a realizar a continuación pero intercambiando los papeles del *IMC* y la edad pero elegimos la variable T como el *IMC* por ser la más usual en el contexto epimediológico.

En el figura 3.1, planteamos una primera estimación empírica de la curva *ROC* y haciendo uso del paquete `pROC`[11] obtenemos tanto el punto de corte óptimo obtenido por el índice de Youden, el *AUC* y un intervalo de confianza para este último. También podemos ver los valores de la especificidad y de la sensibilidad en el punto de corte. De acuerdo con la escala de Swets, el *AUC* obtenido podría ser útil para encontrar una relación entre el *IMC* y la presencia de factores de riesgos cardiovasculares.

A continuación nos planteamos una estimación paramétrica de la curva. Para ello realizamos un test de normalidad por separado para los individuos que tienen riesgo como los que no, obteniendo en ambos casos un p-valor menor que 0.05. Luego se rechaza la normalidad y se plantea una transformación de tipo Box-Cox pero no encontramos ninguna transformación que nos satisfaga.

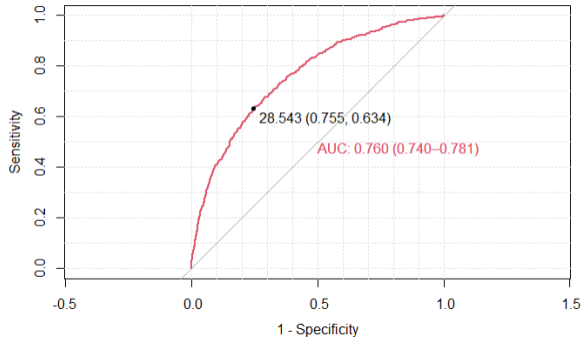


Figura 3.1: Curva *ROC* empírica.

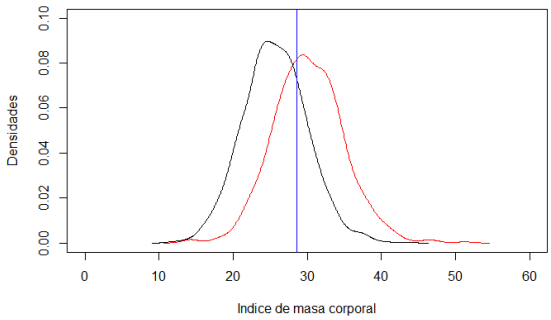


Figura 3.2: Función de densidad y punto de corte óptimo obtenido por el índice de Youden.

Podemos observar en la Figura 3.2, mediante la estimación suavizada de las 2 densidades, el grado de solapamiento de las funciones de densidad no es pequeño, como cabía esperar del valor *AUC* obtenido.

Nos planteamos si el sexo tiene una influencia sobre la curva *ROC*. Realizamos los siguientes pasos para ver si podemos estimar ambas curvas de forma paramétrica:

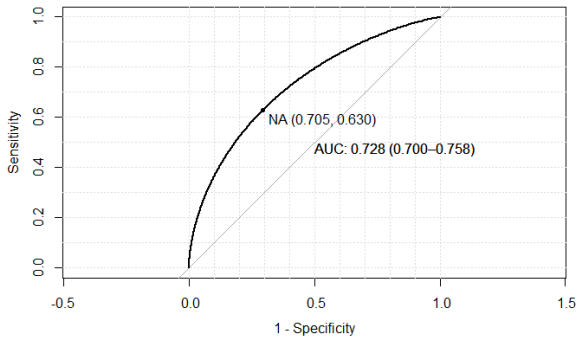
1. Separamos los datos de los hombres y de las mujeres y realizamos un test de normalidad para ambos sexos por separado tanto para individuos con riesgo como no. Los hombres cumplen la normalidad tanto en individuos con riesgo como sin riesgo, con p-valores de 0.27 y 0.39 respectivamente. Para las mujeres obtenemos los p-valores de 0.5 y $0.002 < 0.05$.

2. Planteamos para las mujeres, la transformación de $\sqrt{T}$ y conseguimos la normalidad de nuestros datos que antes no habíamos conseguido, con p-valores 0.35 y 0.89 respectivamente.

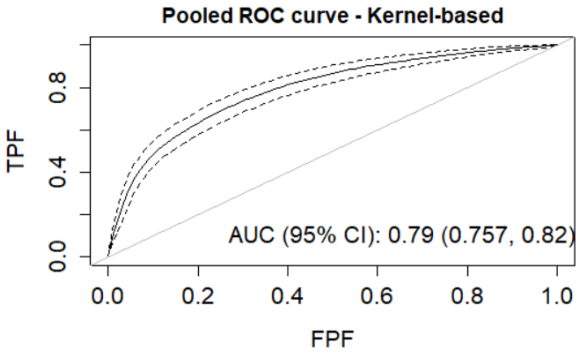
3. Transformar los datos de los hombres y comprobar que se pierde la normalidad ya que se obtienen los p-valores de $0.0007 < 0.05$ para pacientes con riesgo y 0.27 para los que no tienen riesgo.

Así, la curva *ROC* para los hombres la estimamos de forma paramétrica, haciendo uso del paquete `pROC`. Para las mujeres la podríamos representar de manera análoga ya que aunque hemos transformado nuestros datos, sabemos por (1.3), que la curva es invariante frente a transformaciones monótonas. Pero decidimos para las mujeres, representar la curva *ROC* de forma no paramétrica en la que utilizamos la función kernel y el bandwith dados por defecto en el paquete `ROCnReg` [12].

En la figura 3.3*a*, podemos ver los valores de la sensibilidad y especifidad para el punto de corte óptimo pero el paquete `pROC` no nos proporciona dicho punto para el método binormal. Sabemos por (1.20) que podemos calcular el punto de corte óptimo c para el caso binormal, calculándolo obtenemos que $c = 28.05$ para el caso de los hombres. Para la figura 3.3*b*, haciendo uso del paquete `ROCnReg` podemos calcular el punto de corte para la curva de las mujeres, obteniendo $c = 28.45$ y los valores de la sensiblidad y especificidad son 0.69 y 0.74 respectivamente.



(a) Curva *ROC* para hombres por el método Binormal



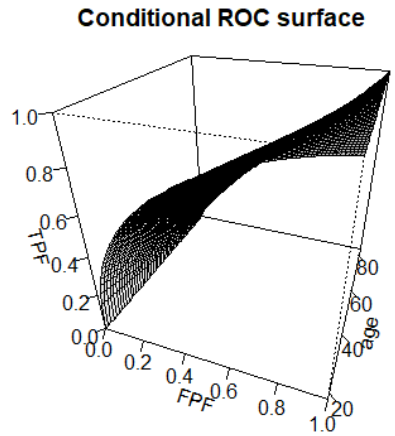
(b) Curva *ROC* para mujeres por el método kernel

Figura 3.3: Curvas *ROC* para ambos sexos por separado

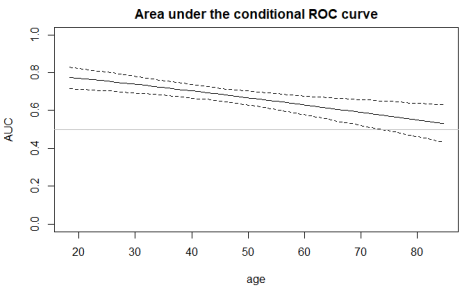
Nos planteamos un test para comparar las curvas vistas hasta ahora utilizando también el paquete `PROC`. Para comparar si el sexo tiene efecto, comparamos las curvas *ROC* estimadas, por ejemplo de forma empírica, de hombres y mujeres. Al obtener un p-valor$= 0.001537 < 0.05$, se deduce que el sexo tiene influencia sobre la curva *ROC* y se debería hacer el estudio por separado para cada sexo.

Como hemos visto que el sexo tiene influencia para estimar la curva *ROC*, vamos a ver si tiene efecto la edad como covariable por separado para los hombres y las mujeres. Primero para los hombres planteamos:

- Haciendo uso del paquete `ROCnReg`, una estimación basada en el método paramétrico.
- Utilizando el paquete `npROCRegression`, una estimación basada en la metodología de la regresión inducida no paramétrica.



(a) Familia de curvas *ROC* dependientes de la edad



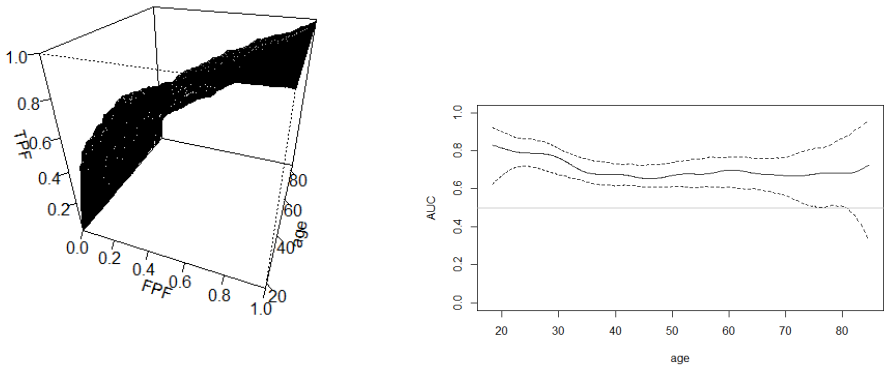
(b) Variación del *AUC* respecto de la edad

Figura 3.4: Curvas *ROC* paramétricas para hombres dependientes de la edad

Estimando con el método paramétrico para los hombres, podemos observar en la figura 3.4*a*, cómo se va aplanando ligeramente la curva conforme va aumentando la edad. Esta idea, se ve reflejada la disminución del *AUC* en la figura 3.4*b*.

Si se plantea una estimación inducida no paramétrica para los hombres, la edad no parece tener prácticamente relevancia respecto la curva, como se muestra en 3.5.

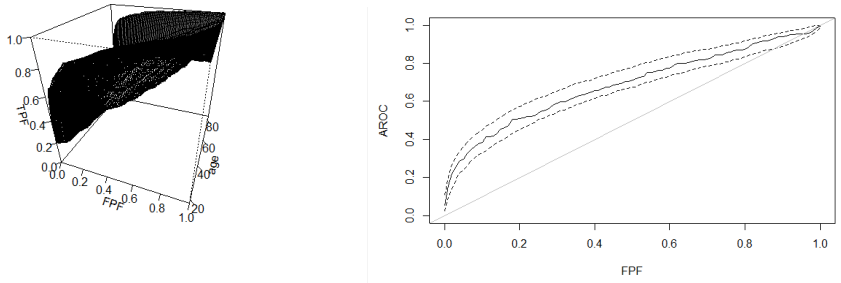
Ahora, utilizando el paquete `npROCRegression`, para las mujeres planteamos lo siguiente:



(a) Familia de curvas *ROC* dependientes de la edad (b) Variación del *AUC* respecto de la edad

Figura 3.5: Curvas *ROC* no paramétricas para hombres dependientes de la edad

- Una estimación inducida no paramétrica del efecto de las covariables.
- Una estimación inducida paramétrica del efecto de las covariables.



(a) Familia de curvas *ROC* dependientes de la edad (b) Curva *ROC* ajustada por la edad

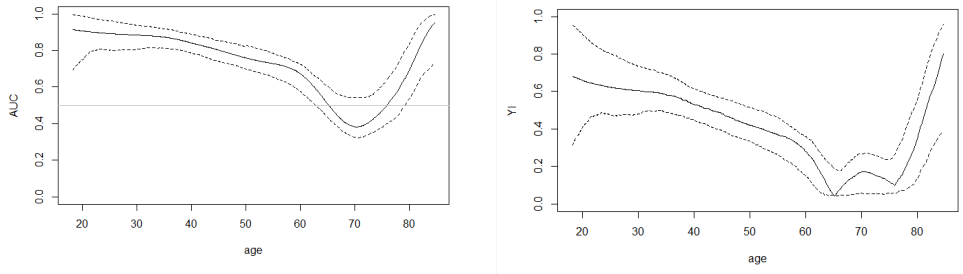
Figura 3.6: Curvas *ROC* para mujeres por estimación inducida no paramétrica

Planteando una estimación inducida no paramétrica para las mujeres, se puede observar en la figura 3.6*a* que para las edades intermedias, la curva *ROC* parece situarse mucho más cerca de la recta $y = x$ que en la gente más jóven o mayor. Esta idea se corrobora en las figuras 3.7*a* y 3.7*b*, donde el *AUC* y el índice de Youden disminuyen para mujeres entre 65 y 75 años.

Si se opta por una estimación inducida paramétrica del efecto de las covariables en el grupo de las mujeres, se obtienen distintos resultados (véase las Figuras 3.7 y 3.8).

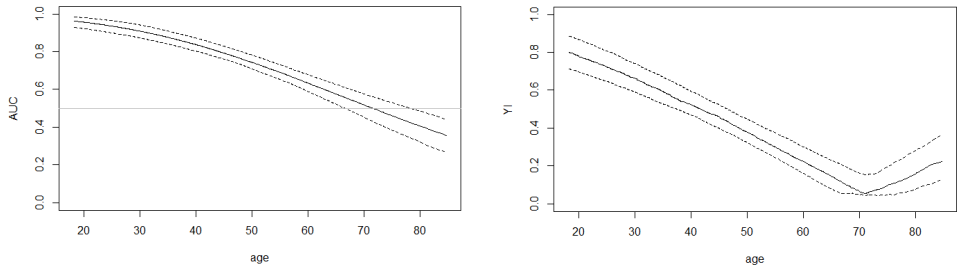
Así, podemos destacar el efecto de la edad en el *AUC* donde el método no paramétrico revela un papel significativo del *IMC* para mujeres mayores de 75 años que el modelo paramétrico no logra identificar.

Concluimos que el sexo tiene influencia sobre la curva *ROC* y que para los hombres la edad parece tener poca influencia sobre la curva *ROC*. Para las mujeres sí que la influencia de la edad sobre la curva *ROC* es más notable y además con una relación no lineal.



(a) Variación del *AUC* respecto de la edad (b) Variación del *YI* respecto de la edad

Figura 3.7: Curvas *ROC* para mujeres por estimación inducida no paramétrica



(a) Variación del *AUC* respecto de la edad (b) Variación del *YI* respecto de la edad

Figura 3.8: Curvas *ROC* para mujeres por estimación inducida paramétrica

Bibliografía

[1] BAMBER, D., *The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph*, Journal of Mathematical Psychology, 12, 387-415, 1975.

[2] CALÌ, C. Y LONGOBARDI, M., *Some mathematical properties of the ROC curve and their applications*, Ricerche di Matematica , 64, 391–402, 2015.

[3] CARNICER ÁLVAREZ, J., *Análisis Numérico*, Apuntes del curso, Departamento de Matemática Aplicada, Universidad de Zaragoza, Zaragoza, 2020.

[4] *Explained: COVID-19 PCR Testing and Cycle Thresholds*, Public Helath Ontario, 17 Feb 2021, Disponible en: `https://www.publichealthontario.ca/en/about/blog/2021/explained-covid19-pcr-testing-and-cycle-thresholds`.

[5] GONCALVES, L. ,SUBTIL, A. , OLIVEIRA, M. Y BERMUDEZ, P., *ROC curve estimation: An overview*, Revstat - Statistical Journal, 12, 1-20, 2014.

[6] HANLEY, J.A. Y MCNEIL, B.J. , *The meaning and use of the area under an ROC curve*, Radiology, 143, 29-36, 1982.

[7] HANLEY, J.A. Y MCNEIL, B.J. *'A method of comparing the areas under receiver operating characteristic curves derived from the same cases'*, Radiology, 148(3), 839-843, 1983.

[8] KRZANOWSKI, W. J. Y HAND, D. J., *ROC Curves for Continuous Data (1st ed.)*, Chapman & Hall/CRC., 2009.

[9] LLOYD, C., *Theory and Methods: Semi-parametric estimation of ROC curves based on binomial regression modelling*, Australian and New Zealand Journal of Statistics,44,75-86, 2002.

[10] PARDO-FERNANDEZ, J. C., RODRIGUEZ-ALVAREZ, M.X. Y VAN KEILGOM, I., *A review on ROC curves in the presence of covariates*, Revstat - Statistical Journal, vol. 12(1), p. 21+, 2014.

[11] ROBIN, X. , TURCK, N. , HAINARD, A. , TIBERTI , N., LISACEK, F., SANCHEZ J. ET AL., *pROC: an open-source package for R and S+ to analyze and compare ROC curves*, BMC Bioinformatics, 12, p 77., 2011, `http://www.biomedcentral.com/1471-2105/12/77/`.

[12] RODRIGUEZ-ALVAREZ, M. X. Y INÁCIO, V., *ROCnReg: ROC Curve Inference with and without Covariates*, R package version 1.0-5, 2021, `https://CRAN.R-project.org/package=ROCnReg`.

[13] RODRIGUEZ-ALVAREZ, M.X. Y ROCA-PARDINAS,J., *npROCRegression: Kernel-Based Nonparametric ROC Regression Modelling*, R package version 1.0-5, 2017, `https://CRAN.R-project.org/package=npROCRegression`.

[14] SILVERMAN, B.W. *Density Estimation in Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[15] VALLE BENAVIDES, A.R.D., *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones*, Trabajo Fin de Grado en Matemáticas, Universidad de Sevilla, Sevilla, 2017.

[16] ZOU, K.H., HALL, W.J. Y SHAPIRO, D.E. *Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests*, Statistics in Medicine, 16(19), 2143-2156, 1997.

Anexo

Escribimos en el anexo el código de R que hemos utilizado para la aplicación a las curvas ROC que hemos realizado. Solo mostramos el código que hemos escrito y no la salida que producen.

Cargamos los paquetes y los datos.

```
library(pROC) # Curvas ROC empíricas y paramétricas sin covariables

## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(npROCRegression)  #Curva ROC dependiente de covariables
de manera no paramétrica
library(ROCnReg) #Curvas ROC no dependientes estimadas con métodos no
# paramétricos y dependientes de covariables estimadadas de forma paramétrica
library(Rcmdr)
library(RcmdrMisc)
library(dplyr)
data("endosim")
```

Realizamos un pequeño análisis descriptivo de las variables de interés

```
numSummary(endosim[,"bmi", drop=FALSE], groups=endosim$idf_status,
  statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))


local({
  .Table <- with(endosim, table(idf_status))
  cat("\ncounts:\n")
  print(.Table)
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
})
numSummary(endosim[,"age", drop=FALSE], groups=endosim$idf_status,
  statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
numSummary(endosim[,"idf_status", drop=FALSE], groups=endosim$gender,
  statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
```

Filtramos por los individos que tienen riesgo y los que no.

```
Riesgo <- filter(endosim,endosim$idf_status==1)
Noriesgo <-filter(endosim,endosim$idf_status==0)
```

Estimamos la curva de forma empírica con el paquete pROC.

```
roc <- roc(endosim$idf_status,endosim$bmi, auc = TRUE, ci = TRUE)
roc
plot.roc( roc, legacy.axes = TRUE, print.thres = "best",
          print.auc = TRUE, auc.polygon = FALSE, max.auc.polygon = FALSE,
          auc.polygon.col = "gainsboro", col = 2, grid = TRUE, ci=TRUE )
auc(roc)
ci(roc)
#coords(roc,transpose = FALSE)
```

Dibujamos la función de densidad y el punto de corte óptimo.

```
dens1 <- density(Riesgo$bmi)
dens2 <- density(Noriesgo$bmi)
plot(NULL,xlim=c(0,60),ylim=c(0,0.1), type="n", xlab="Indice de masa corporal",
     ylab="Densidades")
lines(dens1,col="red")
lines(dens2)
abline(v=28.5427,col="blue")
```

Vemos si nuestros datos pasan el test de normalidad.

```
normalityTest(~bmi, test="shapiro.test", data=Riesgo)

normalityTest(~bmi, test="shapiro.test", data=Noriesgo)
```

Como no son normales, proponemos una posible transformación de nuestros datos y deberiamos elevar nuestros datos a 0.73 lo que no tiene sentido matemático.

```
summary(powerTransform(bmi ~ 1, data=Noriesgo, family="bcPower"))
summary(powerTransform(bmi ~ 1, data=Riesgo, family="bcPower"))
```

Ahora filtramos por sexo para ver si tiene efecto sobre la curva ROC. Los sometemos a un test de normalidad.

```
Chicas <- filter(endosim,endosim$gender== "Women")
Chicos <- filter(endosim,endosim$gender== "Men")

Riesgochicas <- filter(Chicas, Chicas$idf_status== 1)
normalityTest(~bmi, test="shapiro.test", data=Riesgochicas)

Noriesgochicas <-filter(Chicas,Chicas$idf_status== 0)
normalityTest(~bmi, test="shapiro.test", data=Noriesgochicas)

Riesgochicos <- filter(Chicos, Chicos$idf_status== 1)
normalityTest(~bmi, test="shapiro.test", data=Riesgochicos)

Noriesgochicos <-filter(Chicos,Chicos$idf_status== 0)
normalityTest(~bmi, test="shapiro.test", data=Noriesgochicos)
```

Como los hombres lo pasan, planteamos una transfromación en las mujeres.

```
summary(powerTransform(bmi ~ 1, data=Riesgochicas, family="bcPower"))
Riesgochicas$sqrtbmi <- sqrt(Riesgochicas$bmi)
normalityTest(~sqrtbmi, test="shapiro.test", data=Riesgochicas)
summary(powerTransform(bmi ~ 1, data=Noriesgochicas, family="bcPower"))
Noriesgochicas$sqrtbmi <- sqrt(Noriesgochicas$bmi)
normalityTest(~sqrtbmi, test="shapiro.test", data=Noriesgochicas)
```

Si se transforma a la raiz de los datos, la normalidad se consigue para mujeres pero se pierde para los hombres.

```
Riesgochicos$sqrtbmi <- sqrt(Riesgochicos$bmi)
normalityTest(~sqrtbmi, test="shapiro.test", data=Riesgochicos)
Noriesgochicos$sqrtbmi <- sqrt(Noriesgochicos$bmi)
normalityTest(~sqrtbmi, test="shapiro.test", data=Noriesgochicos)
```

Realizamos la curva ROC paramétrica para los hombres.

```
rocchicos1 <- roc(Chicos$idf_status,Chicos$bmi, auc = TRUE, ci = TRUE,
                  smooth=TRUE, smooth.method="binormal")
rocchicos1
auc(rocchicos1)
ci(rocchicos1)
plot.roc( rocchicos1, legacy.axes = TRUE,  print.thres = "best" ,print.auc = TRUE,
          auc.polygon = FALSE, max.auc.polygon = FALSE, auc.polygon.col = "gainsboro",
          col = 2, grid = TRUE )
```

Calculamos el punto de corte

```
mean(Riesgochicos$bmi)
sd(Riesgochicos$bmi)
mean(Noriesgochicos$bmi)
sd(Noriesgochicos$bmi)
```

```
c <- function(a,b,c,d){

  c*b^2-a*d^2-b*d*(sqrt((c-a)^2+(b^2-d^2)*log(b^2/d^2)))/(b^2-d^2)

}
```

```
c(26.01731,3.71665,29.36563,4.099683)
```

Dibujamos para las mujeres una curva *ROC* basada en el método kernel con el paquete `ROCnReg`.

```
m0_kernel <- pooledROC.kernel(marker = "bmi", group = "idf_status",
tag.h = 0, data = Chicas, p = seq(0,1,l=101), bw ="SRT",
B = 500, method = "coutcome",
pauc = pauccontrol(compute = TRUE, value = 0.5, focus = "FPF"))
```

Calculamos el punto de corte y los valores para la sensibilidad y especifidad en ese punto.

```
summary(m0_kernel)
plot(m0_kernel)
th_m0_yi <- compute.threshold.pooledROC(m0_kernel, criterion = "YI")
th_m0_yi$threshold
1-th_m0_yi$FPF
th_m0_yi$TPF
```

Test de hipótesis para las curvas empíricas de hombres y mujeres.

```
rocchicas <- roc(Chicas$idf_status,Chicas$bmi, auc = TRUE, ci = TRUE)
rocchicos <- roc(Chicos$idf_status,Chicos$bmi, auc = TRUE, ci = TRUE)
roc.test(rocchicas,rocchicos)
```

Curvas ROC paramétrica para hombres dependiente de la edad con el paquete `ROCnReg`.

```
agep <- seq(min(Chicos$age), max(Chicos$age), length = 50)
df.pred <- data.frame(age = agep)

cROC_sp_normal <- cROC.sp(formula.h = bmi ~ age,
                          formula.d = bmi ~ age,
                          group = "idf_status",
                          tag.h = 0,
                          data = Chicos ,
                          est.cdf = "normal",
                          newdata = df.pred,
                          pauc = list(compute = TRUE, value = 0.5, focus = "FPF"),
                          p = seq(0, 1, l = 101),
                          B = 1000)
summary(cROC_sp_normal)

plot(cROC_sp_normal)
```

Curva ROC suavizada no paramétrica para hombres dependiente de la edad.

```
m0.men <- INPROCreg(marker = "bmi", covariate = "age", group = "idf_status",
                    tag.healthy = 0,
                    data = Chicos,
                    ci.fit = TRUE, test = TRUE,
                    accuracy = c("YI","TH"),
                    accuracy.cal=c("ROC","AROC"),
                    control=controlINPROCreg(p=1,kbin=30,step.p=0.01),
                    newdata = NULL)
m0.men
summary(m0.men)
plot(m0.men)
```

Curva ROC suavizada no paramétrica para mujeres dependiente de la edad con el paquete `npROCRegression`.

```
m0.women <- INPROCreg(marker = "bmi", covariate = "age", group = "idf_status",
                    tag.healthy = 0,
```

```
                    data = Chicas,
                    ci.fit = TRUE, test = TRUE,
                    accuracy = c("YI","TH"),
                    accuracy.cal=c("ROC","AROC"),
                    control=controlINPROCreg(p=1,kbin=30,step.p=0.01),
                    newdata = NULL)
summary(m0.women)
plot(m0.women)
```

Curva ROC suavizada no paramétrica para mujeres dependiente de la edad con el paquete `npROCRegression` pero estimando las covariables por métodos paramétricos.

```
m1.women <- INPROCreg(marker = "bmi", covariate = "age", group = "idf_status",
                    tag.healthy = 0,
                    data = Chicas,
                    ci.fit = TRUE, test = TRUE,
                    accuracy = c("YI","TH"),
                    accuracy.cal=c("ROC","AROC"),
                    control=controlINPROCreg(p=1,kbin=30,step.p=0.01,h=c(0,0,0,0)),
                    newdata = NULL)
summary(m1.women)
plot(m1.women)
```