UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Sensibilidad y estimación robusta en modelos de regresión directa para curvas ROC

Jésica Charaf

Directora: Dra. Ana M. Bianco

Junio de 2022

Agradecimientos

A Ana Bianco por dirigir esta tesis y acompañarme durante todo este último tiempo de mi carrera con tanta dedicación y esfuerzo, en un contexto muy particular. Por ser una excelente directora y una gran docente, que transmite sus conocimientos con deseo y de quien aprendí muchísimo. Por toda su paciencia y empatía. Por hacerme devoluciones detalladas y ser tan precisa, por motivarme, intercambiar conmigo y escuchar cada una de mis inquietudes.

A Mariela Sued y Marina Valdora por haber aceptado la invitación a leer esta tesis y conformar el jurado.

Agradezco a la educación pública, a la Universidad de Buenos Aires y a la FCEyN por brindarme la posibilidad de tener una formación de calidad.

A mis docentes por todo lo que me enseñaron, los desafíos, los espacios de consultas e intercambio y por toda su vocación y compromiso.

A mis amigas y amigos que son fundamentales, por estar siempre apoyándome, hacerme reír y despejarme en cualquier momento y por compartir la vida conmigo. A mis compañeros y compañeras de la carrera, del trabajo y del equipo de apoyo escolar, por todos los intercambios y reflexiones compartidas que conforman el aprendizaje, por la construcción conjunta y por hacer que el camino sea más llevadero. A Meli Scotti quien estuvo presente a lo largo de toda mi carrera y en cada momento importante, sacándome siempre una sonrisa. Y muy especialmente a Anita y a Nati, por estar junto a mí este último tiempo aguantándome y ayudándome, porque realmente fueron una parte importantísima de todo este proceso, por leerme y escuchar mis infinitos audios y por involucrarse en esta tesis con tanto amor.

Gracias a mi familia, soporte indispensable, por apoyarme en todo, interesarse en mi carrera y confiar siempre en mí. A mi mamá y a mi papá por aguantarme infinitamente, escucharme y estar siempre, por todas las oportunidades que me dieron, por incentivarme a estudiar y darme la posibilidad de elegir y por apoyar todas mis decisiones. A Sabri y a Darío, mis hermanos y amigos incondicionales, por ser mis guías y referentes, por tener la suerte de haber crecido junto a ellos, por enseñarme tanto y siempre tener la palabra justa, por hacerme saber que nunca voy a estar sola. A Simón por ser una personita maravillosa y alegrarme cada día.

A Mati, mi sostén fundamental y compañero de vida, por apoyarme, incentivarme y creer en mí. Por ser partícipe de esta tesis con cuerpo y alma. Por toda su paciencia, las horas de escucha y la inmensa ayuda. Por intercambiar, reflexionar y leer la tesis conmigo. Por esperarme hasta cualquier hora. Y, por sobre todas las cosas, por estar junto a mí estos años y por todo su amor.

Índice general

Introducción

Las curvas ROC (Receiver Operating Characteristic) son una herramienta útil para evaluar el desempeño de una variable continua o de una prueba diagnóstica en la clasificación entre dos estados. Fueron desarrolladas en el contexto de la Teoría de Detección de Señales durante la Segunda Guerra Mundial para analizar la capacidad de un operador de distinguir un objeto enemigo en la pantalla de un radar. Las curvas ROC se expandieron a distintos campos y, en particular, se hicieron muy populares en el área de la Medicina a partir de los años 1960. Tienen un gran potencial en estudios médicos que analizan el desempeño de pruebas y biomarcadores para diagnosticar una enfermedad.

En el área de Estadística las curvas ROC se utilizan ampliamente en situaciones de clasificación y discriminación, donde típicamente se tienen dos estados o clases y, a partir de un clasificador binario, se asigna a los individuos de un conjunto uno de los estados según un valor de corte. La curva ROC es una representación gráfica que permite visualizar la capacidad discriminatoria de un clasificador binario a medida que varía el valor crítico de decisión.

Para ejemplificar fijaremos nuestro escenario en el campo del diagnóstico médico. Asumamos que tenemos un clasificador para distinguir entre dos poblaciones, enfermos y sanos, basado en la medición de un biomarcador o una variable continua, a quien llamaremos Y, y un valor de umbral c que determinará la regla de asignación. Supongamos, entonces, que un individuo es clasificado como enfermo si Y es mayor o igual a c y como sano, en caso contrario. En dicha clasificación se pueden cometer errores; por ejemplo, podría ocurrir que un individuo sano arroje un valor de Y mayor o igual a c y sea asignado incorrectamente a la clase enferma, incurriendo en lo que llamamos un falso positivo. En contraposición, los verdaderos positivos corresponden a los individuos enfermos que son clasificados correctamente. Es claro que según cuál sea el valor de decisión c que tomemos, se presentará una variación en la proporción de falsos positivos y verdaderos positivos. La curva ROC representa gráficamente la relación entre la proporción de falsos positivos en la población sana y la proporción de verdaderos positivos en la población enferma

a medida que varía el valor de corte. En este sentido es que la curva ROC resulta un instrumento de interés para evaluar la eficacia de un clasificador o, a su vez, para comparar distintos procedimientos de decisión.

En muchas situaciones se dispone de variables con información adicional de cada individuo de la población. Como se menciona en Pardo-Fernández *et al.* (2014), cuando hay presencia de covariables es recomendable incorporarlas en el análisis de la curva ROC ya que la efectividad y la capacidad discriminatoria de la prueba diagnóstica podrían verse afectadas por las mismas. En este sentido, Pepe (2003) muestra diferentes circunstancias en las que las covariables impactan en el resultado de un test o biomarcador. Por ejemplo, en el caso del diagnóstico de un paciente, caracterísiticas tales como la edad o el sexo pueden ser de relevancia. La incoporación de esa información adicional se puede realizar a partir de la obtención de curvas ROC condicionadas a valores específicos de las covariables y, en muchas ocasiones, permite tener una mayor comprensión de la efectividad del test que si se analizara la curva ROC ignorando esa distinción.

Existen diversos métodos propuestos en la literatura para incorporar la información de las covariables a las curvas ROC; entre ellos se encuentran la metodología inducida y la metodología directa. En la metodología inducida se ajusta un modelo de regresión por separado al biomarcador en cada una de las poblaciones en términos de las covariables, y luego se deriva la curva ROC inducida. En cambio, en la metodología directa se asume un modelo de regresión lineal generalizado para la curva ROC y los efectos de las covariables en la misma se evalúan directamente. En ambas metodologías se han propuesto distintas técnicas para estimar las curvas ROC de forma paramétrica, semiparamétrica y no paramétrica.

En el presente trabajo nos interesa analizar la robustez de la estimación de las curvas ROC, en el sentido de cuán resistente resulta ante pequeñas desviaciones del modelo asumido o en presencia de un porcentaje moderado de datos atípicos. Rodríguez-Álvarez *et al.* (2011) estudian la estabilidad de los estimadores de las curvas ROC frente al modelado erróneo del efecto de las covariables para diversos modelos y, en particular para la metodología directa, analizan el impacto que tiene la especificación incorrecta del modelo de regresión utilizado para la curva ROC. Más recientemente, Bianco, Boente y González-Manteiga (2020) analizan la robustez de la estimación de la curva ROC dentro de la metodología inducida y exploran estimadores robustos que se muestren resistentes ante la presencia de un porcentaje de datos atípicos.

Teniendo todo esto en cuenta, estudiaremos la estimación de curvas ROC en presencia de covariables y enmarcaremos nuestro trabajo dentro de la metodología directa. Nos proponemos indagar sobre la sensibilidad de estos modelos ante la presencia de un porcentaje de datos atípicos y en qué medida resisten a conta-

minaciones. A su vez, exploraremos distintas herramientas de estimación robusta e implementaremos modificaciones en las distintas etapas del método clásico de estimación con el objetivo de obtener métodos resistentes ante pequeñas desviaciones del modelo asumido y que, además, resulten eficientes cuando el modelo se sostiene.

El trabajo de tesis está organizado de la siguiente manera. En el Capítulo 1 haremos una introducción conceptual de las curvas ROC, las definiciones básicas y sus propiedades. En el Capítulo 2 revisaremos la metodología directa para estimar curvas ROC en presencia de covariables y los métodos de estimación involucrados en este procedimiento. En el Capítulo 3 abordaremos distintas herramientas de estimación robusta, tanto en el marco paramétrico como en el no paramétrico, e introduciremos una propuesta robusta para la estimación de la curva ROC en la que se combinan varios de los métodos recorridos. En el Capítulo 4 presentaremos un estudio de simulación numérico en donde analizamos la sensibilidad de los estimadores clásicos y de los distintos estimadores propuestos para la curva ROC. Por último, en el Capítulo 5 aplicaremos las metodologías desarrolladas en un conjunto de datos reales.

Capítulo 1

Curvas ROC

En este capítulo nos concentraremos en la definición formal de las curvas ROC e introduciremos algunas nociones y propiedades básicas.

1.1. Definición de la curva ROC

Vamos a asumir que trabajamos con dos poblaciones, donde los individuos de cada una de ellas serán identificados como enfermos y sanos, y que disponemos de una variable continua de diagnóstico Y, también llamada biomarcador, que se considera para la asignación de un individuo de la población en alguna de las dos clases a partir de un valor de corte c. Supondremos que dado un valor crítico c un individuo será clasificado como:

- enfermo si $Y \geq c$,
- sano si $Y < c$.

Sea F_D la distribución del biomarcador Y en la población enferma y F_H la distribución del mismo en la población sana. De ahora en adelante, notaremos con $Y_D \sim F_D$ al biomarcador en la población de enfermos y con $Y_H \sim F_H$ al biomarcador en la población de sanos. Sin pérdida de generalidad, asumiremos que Y_D es estocásticamente mayor que Y_H, es decir, que $F_D(c) \leq F_H(c)$ para todo c.

Para cualquier valor de corte c que se elija, es posible que se presenten errores en la clasificación. Por ejemplo, podría suceder que un individuo sano arroje un valor $Y_H \geq c$ y sea clasificado como enfermo, conduciendo a lo que llamamos

un falso positivo. A su vez, puede ocurrir que un individuo enfermo presente un valor $Y_D < c$, lo que correspondería a un falso negativo. En este sentido, se definen la *fracción de verdaderos positivos* (TPF), que representa la probabilidad de que un individuo enfermo sea diagnosticado como tal, y la *fracción de falsos positivos* (FPF), correspondiente a la probabilidad de que un individuo sano sea diagnosticado como enfermo:

$$\begin{aligned} \text{TPF}(c) &= P(Y_D \geq c) \\ \text{FPF}(c) &= P(Y_H \geq c)\,. \end{aligned}$$

La *sensibilidad* de un test o biomarcador se define como $Se(c) = 1 - F_D(c)$ y representa a la probabilidad de clasificar a un individuo enfermo como enfermo, es decir, $Se(c) = \text{TPF}(c)$. Por otra parte, la *especificidad* viene dada por $Sp(c) = F_H(c)$ y corresponde a la probabilidad de clasificar a un individuo sano como sano. De este modo, se tiene que $\text{FPF}(c) = 1 - Sp(c) = 1 - F_H(c)$.

En la Figura 1.1 observamos un ejemplo de la representación gráfica de las fracciones $\text{TPF}(c)$ y $\text{FPF}(c)$ para un cierto valor fijo del umbral c. Por arriba del eje de las abscisas está representada la distribución del test diagnóstico en la población enferma y, por debajo, la distribución del test en la población sana. El área sombreada en violeta a la derecha del valor c corresponde a $\text{TPF}(c)$ y el área sombreada en rojo corresponde a $\text{FPF}(c)$. Como podemos observar los errores de clasificación dependerán del valor de c.

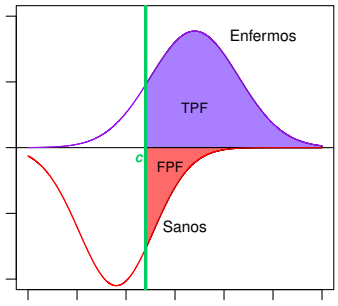


Figura 1.1: Distribución del test en la población enferma y en la población sana (arriba y debajo del eje x, respectivamente). La línea vertical verde corresponde al umbral c. En violeta está sombreada el área correspondiente a TPF y en rojo el área correpondiente a FPF.

Las ternas $\{(c, \text{FPF}(c), \text{TPF}(c)), \quad c \in \mathbb{R}\}$ constituyen un objeto geométrico llamado *curva ROC* y resultan de interés dado que permiten representar la relación entre la proporción de falsos positivos y verdaderos positivos en un test, a medida que varía el umbral c. En este sentido, la curva ROC es una herramienta para evaluar la capacidad discriminatoria del biomarcador. A su vez, podemos describir a la curva ROC a partir de los pares

$$\{(\text{FPF}(c), \text{TPF}(c)), \quad c \in (-\infty, \infty)\} =$$

$$\{(1 - F_H(c), 1 - F_D(c)), \quad c \in (-\infty, \infty)\}.$$

Si realizamos una reparametrización de la curva, llamando $t = 1 - F_H(c)$, obtenemos el conjunto de pares:

$$\left\{\left(t, 1 - F_D(F_H^{-1}(1-t))\right), \quad t \in [0,1]\right\}.$$

Esto nos permite escribir a la curva ROC de la siguiente forma:

$$\text{ROC}(t) = 1 - F_D(F_H^{-1}(1-t)), \quad 0 \leq t \leq 1\,. \tag{1.1}$$

Observemos que, en base a las propiedades de toda función de distribución acumulada, la curva ROC resulta monótona creciente puesto que $\text{FPF}(c)$ y $\text{TPF}(c)$ decrecen tendiendo a 0 a medida que c aumenta y ambas crecen aproximándose a 1 cuando c tiende a $-\infty$.

1.2. Propiedades elementales de las curvas ROC

A continuación mencionaremos algunas características y propiedades básicas de las curvas ROC que serán de utilidad a lo largo de este trabajo, tomando como referencia resultados que se encuentran desarrollados en Pepe (2003).

Como vimos previamente, la curva ROC es una función monótona creciente que mapea el intervalo $[0,1]$ en el $[0,1]$. Además, tiene la propiedad de ser invariante frente a cualquier transformación monótona creciente de la variable Y, lo cual demostraremos a partir de la siguiente proposición.

Proposición 1.2.1. *Sean Y una variable aleatoria y la curva ROC definida a partir de los pares $\{(FPF(c), TPF(c)), \quad c \in (-\infty, \infty)\}$, entonces se tiene que la curva ROC resulta invariante ante transformaciones monótonas crecientes de la variable Y.*

Demostración. Sea h una transformación monótona creciente y consideremos la variable $W = h(Y)$. Notaremos con $W_D = h(Y_D)$ y $W_H = h(Y_H)$ a la variable W en la población enferma y sana, respectivamente. Veamos que cualquier punto $(\text{FPF}(c), \text{TPF}(c))$ perteneciente a la curva ROC de Y también es un punto que pertenece a la curva ROC asociada a W.

En efecto, si consideramos $d = h(c)$ tenemos que

$$P(W_H \geq d) = P(h(Y_H) \geq h(c)) = P(Y_H \geq c)$$

y

$$P(W_D \geq d) = P(h(Y_D) \geq h(c)) = P(Y_D \geq c).$$

Luego, el par $(\text{FPF}(c), \text{TPF}(c))$ corresponde también a un punto perteneciente a la curva ROC para W.

De forma análoga, se puede probar que cualquier punto perteneciente a la curva ROC asociada a W se encuentra en la curva ROC de la variable Y.

□

Una vez definidas las curvas ROC, resulta de interés analizar cuáles son las características de las mismas que nos permiten distinguir entre un test con mayor o menor capacidad discriminatoria. Como se menciona en Pepe (2003), un biomarcador *no informativo* es aquel en el cual la variable Y no tiene relación con el estado de la enfermedad de un sujeto, mientras que un test *perfecto* separa completamente a los sujetos sanos de los enfermos. En el caso de un biomarcador no informativo se tiene que las distribuciones de Y en la población enferma y en la sana (es decir, Y_D e Y_H) son iguales, por lo que $\text{TPF}(c) = \text{FPF}(c) = t$ para cualquier valor de c y la curva ROC resulta ser $\text{ROC}(t) = t$. Por otra parte, para un test perfecto existe un valor de c tal que $\text{TPF}(c) = 1$ y $\text{FPF}(c) = 0$, lo que conduce a una curva ROC asociada que se encuentra a lo largo de los bordes izquierdo y superior del cuadrado unitario del primer cuadrante. En la Figura 1.2 se representan en paralelo distintas curvas ROC junto a las gráficas de las correspondientes distribuciones de Y_D e Y_H, entre las cuales se encuentran las curvas asociadas a un biomarcador no informativo y a uno perfecto.

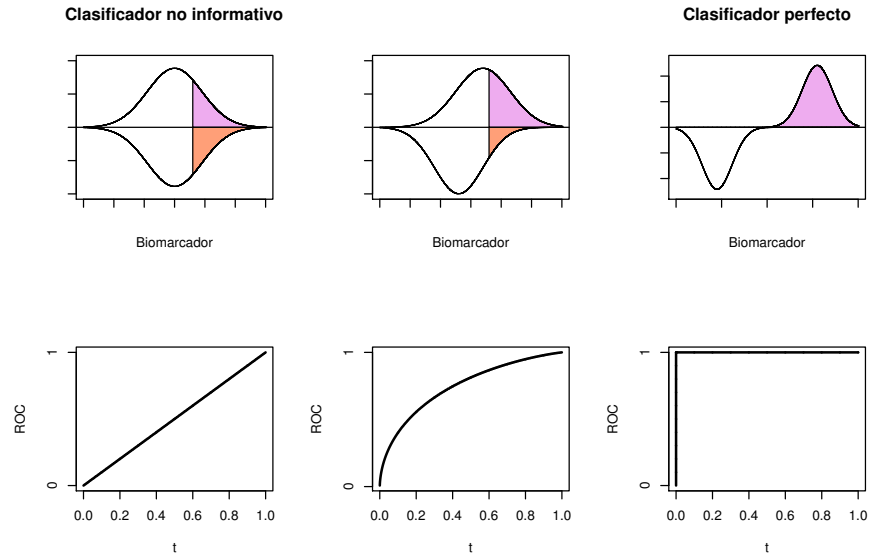


Figura 1.2: Curvas ROC asociadas a diferentes distribuciones F_H y F_D. En el panel izquierdo se muestra la curva ROC de un biomarcador no informativo y en el derecho la de uno perfecto.

Los tests que tienen mejor capacidad discriminatoria son aquellos cuyas curvas ROC asociadas son más parecidas a la curva perfecta, es decir, más cercanas a la esquina superior izquierda del cuadrado unitario. Este hecho lo podemos observar en la Figura 1.3 en la cual se encuentran representadas las curvas ROC asociadas a dos tests distintos A y B. Para cualquier fracción de falsos positivos que se considere, la fracción de verdaderos positivos será mayor en el test A que en el B y de allí podemos deducir que el test A presenta un mejor desempeño. Análogamente, si consideramos umbrales tales que el valor de TPF sea el mismo para ambos biomarcadores, obtenemos que la fracción de falsos positivos es menor en el test A que en el B.

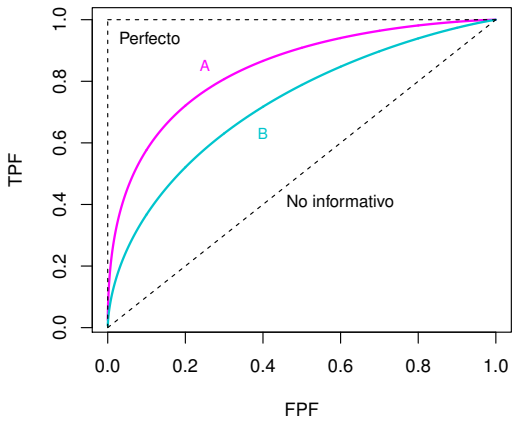


Figura 1.3: Curvas ROC para dos tests A y B, donde A es mejor que B.

1.3. Medidas de resumen

En la literatura existen diversas propuestas de medidas de resumen de la curva ROC que son consideradas para evaluar la capacidad discriminatoria de un test. Dichas medidas consisten en índices numéricos que transmiten información importante de la curva y permiten caracterizarla. En particular, resultan de utilidad al momento de comparar la efectividad de distintos biomarcadores. Nos referiremos a Pepe (2003) para un recorrido por las distintas medidas de resumen que más se utilizan, donde se pueden encontrar más detalles. Una de las medidas más populares es el área bajo la curva ROC.

1.3.1. Área bajo la curva (AUC)

El área bajo la curva ROC, que notaremos como AUC por sus siglas en inglés (Area Under the Curve), es uno de los índices más usados y se define de la siguiente manera:

$$\text{AUC} = \int_0^1 \text{ROC}(t)dt\,.$$

En la Figura 1.4 se presentan distintos ejemplos de curvas ROC y sus respectivas AUC, entre las cuales se encuentran las asociadas a un test perfecto y a uno no informativo.

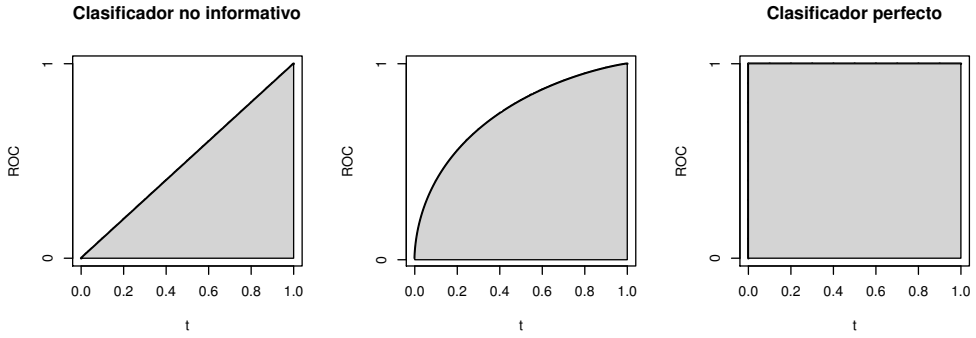


Figura 1.4: Área bajo la curva ROC para distintos biomarcadores.

Como podemos observar, el área bajo la curva correspondiente a un test perfecto es igual a 1, mientras que un test no informativo tiene AUC = 0.5. Luego, es esperable que un biomarcador con mayor precisión en la clasificación tenga un valor de AUC más cercano a 1 que un biomarcador de peor capacidad discriminatoria, puesto que su curva ROC estará más próxima a la de un test perfecto.

Es claro que si A es un test uniformemente mayor que B, es decir, que verifica que $\mathrm{ROC}_A(t) \geq \mathrm{ROC}_B(t)\ \forall t \in [0,1]$, entonces sus AUC también estarán ordenadas de modo que $\mathrm{AUC}_A \geq \mathrm{AUC}_B$. La afirmación contrarrecíproca no necesariamente es válida.

El AUC puede interpretarse como la probabilidad de que el valor del test resulte mayor en un sujeto enfermo que en uno sano, esto es, $\mathrm{AUC} = P(Y_D > Y_H)$. A continuación daremos la demostración formal de esta propiedad.

Proposición 1.3.1. *Si $Y_D \sim F_D$ e $Y_H \sim F_H$ son variables aleatorias independientes que corresponden al resultado de un test en las poblaciones enferma y sana, respectivamente, y f_H es la función de densidad de Y_H, entonces:*

$$AUC = P(Y_D > Y_H)\,.$$

Demostración. Tenemos que:

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t)dt = \int_0^1 \left[1 - F_D(F_H^{-1}(1-t))\right] dt\,.$$

Usando el cambio de variables $y = F_H^{-1}(1-t)$ podemos escribir:

$$\mathrm{AUC} = \int_{+\infty}^{-\infty} \left[1 - F_D(y)\right] d(1 - F_H(y)) = \int_{-\infty}^{+\infty} P(Y_D > y) f_H(y)\, dy\,,$$

donde $f_H(y)$ denota la función de densidad correspondiente a la distribución de Y_H.

Luego, como Y_H e Y_D son independientes, se tiene que:

$$\text{AUC} = \int_{-\infty}^{+\infty} P(Y_D > y, Y_H = y)\, dy = P(Y_D > Y_H)\,.$$

□

Este hecho refuerza la idea de que valores de AUC cercanos a 1 sugieren una capacidad de diagnóstico alta del biomarcador, ya que dicha medida representa la probabilidad de que los resultados del test en un sujeto enfermo y en uno sano estén correctamente ordenados.

Otra forma posible de interpretar el AUC es como un promedio de los valores de TPF, tomado uniformemente sobre todo el rango de fracciones de falsos positivos en el intervalo $(0, 1)$.

1.3.2. Otras medidas de resumen

Además del AUC, que es uno de los índices que más frecuentemente se utilizan, se han propuesto diversas medidas para resumir información de la curva ROC.

En algunas ocasiones resulta de interés analizar un valor específico de fracción de falsos positivos, t_0, y su valor correspondiente de TPF provee información relevante. Con lo cual, se puede considerar como medida de resumen un punto ROC específico, es decir:

$$\text{ROC}(t_0).$$

Por ejemplo, en casos en los que es posible acomodar la fracción de falsos positivos a un valor t_0 aceptable, se pueden observar los respectivos resultados de $\text{ROC}(t_0)$ para comparar dos tests que producen ese valor de FPF.

Por otra parte, hay situciones en las que puede ser de utilidad analizar los valores de $\text{ROC}(t)$ para $t < t_0$. A diferencia de la medida $\text{ROC}(t_0)$, que ignora gran parte de la información de la curva ROC, el área bajo la curva parcial es un índice que restringe la atención a fracciones de falsos positivos menores o iguales a t_0, utilizando todos los puntos de la curva ROC en ese rango. El área bajo la curva parcial, $\text{pAUC}(t_0)$, se define como:

$$\text{pAUC}(t_0) = \int_0^{t_0} \text{ROC}(t) dt\,.$$

La medida de Kolmogorov-Smirnov (KS) también es muy utilizada y representa la distancia vertical máxima que hay entre la curva ROC y la recta asociada a la función identidad, indicando cuán lejos se encuentra la curva de un test no informativo. Está definida como:

$$\mathrm{KS} = \max_t |\mathrm{ROC}(t) - t| \, .$$

Los valores varían en el rango $[0, 1]$, siendo 0 el índice para un test no informativo y 1 para el test perfecto.

La medida KS puede verse como una generalización del Índice de Youden, que para tests binarios se define como TPF − FPF y, en general, viene dado por:

$$\max_c \{Se(c) + Sp(c) - 1\} \, .$$

Otra medida que se considera de interés es la que en Pepe (2003) se denomina como "punto de simetría" y se denota con Sym. Esta medida indica el punto de la curva ROC donde la sensibilidad es igual a la especificidad, $\mathrm{TPF} = 1 - \mathrm{FPF}$, y queda definida por

$$\mathrm{ROC}(\mathrm{Sym}) = 1 - \mathrm{Sym}.$$

1.4. Modelo ROC binormal

El modelo ROC binormal es uno de los más comúnmente utilizados y tiene un rol central en el análisis de curvas ROC. Es aplicable cuando el biomarcador sigue una distribución normal tanto en la población enferma como en la sana. A partir de este supuesto, se puede deducir la forma funcional de la curva ROC.

Proposición 1.4.1. *Si* $Y_H \sim N(\mu_H, \sigma_H^2)$, $Y_D \sim N(\mu_D, \sigma_D^2)$, *luego*

$$ROC(t) = \phi(a + b\phi^{-1}(t)), \tag{1.2}$$

donde

$$a = \frac{\mu_D - \mu_H}{\sigma_D} \tag{1.3}$$

$$b = \frac{\sigma_H}{\sigma_D} \tag{1.4}$$

y ϕ *denota la función de distribución acumulada de una variable aleatoria normal estándar.*

Demostración. Para cualquier valor de umbral c, usando la simetría respecto de cero se tiene que:

$$\begin{aligned} \text{FPF}(c) &= P(Y_H \geq c) = \phi\left(\frac{\mu_H - c}{\sigma_H}\right), \\ \text{TPF}(c) &= P(Y_D \geq c) = \phi\left(\frac{\mu_D - c}{\sigma_D}\right). \end{aligned}$$

Tomando cualquier fracción de falsos positivos $t = \text{FPF}(c)$, el umbral correspondiente resulta ser $c = \mu_H - \sigma_H \phi^{-1}(t)$. Luego:

$$\begin{aligned} \text{ROC}(t) = \text{TPF}(c) &= \phi\left(\frac{\mu_D - c}{\sigma_D}\right) \\ &= \phi\left(\frac{\mu_D - \mu_H + \sigma_H \phi^{-1}(t)}{\sigma_D}\right) \\ &= \phi(a + b\phi^{-1}(t)). \end{aligned}$$

□

A su vez, en el modelo ROC binormal es posible determinar una fórmula cerrada para el AUC como se muestra en la siguiente proposición.

Proposición 1.4.2. *Supongamos $Y_H \sim N(\mu_H, \sigma_H^2)$ e $Y_D \sim N(\mu_D, \sigma_D^2)$, variables aleatorias independientes, entonces*

$$AUC = \phi\left(\frac{a}{\sqrt{1+b^2}}\right), \tag{1.5}$$

con a y b definidos en (1.3) *y* (1.4).

Demostración. Recordemos que $\text{AUC} = P(Y_D > Y_H) = P(Y_D - Y_H > 0)$. Sea $W = Y_D - Y_H$. Luego, como Y_D e Y_H son independientes,

$$W \sim N(\mu_D - \mu_H, \sigma_D^2 + \sigma_H^2).$$

Por lo tanto,

$$\begin{aligned} P(W > 0) &= 1 - \phi\left(\frac{-\mu_D + \mu_H}{\sqrt{\sigma_D^2 + \sigma_H^2}}\right) = \phi\left(\frac{\mu_D - \mu_H}{\sqrt{\sigma_D^2 + \sigma_H^2}}\right) \\ &= \phi\left(\frac{\mu_D - \mu_H}{\sigma_D} \Big/ \sqrt{1 + \frac{\sigma_H^2}{\sigma_D^2}}\right) \\ &= \phi\left(\frac{a}{\sqrt{1+b^2}}\right) \end{aligned}$$

□

En la Figura 1.5 se muestra un ejemplo de la curva ROC para un modelo binormal donde $Y_H \sim N(\mu_H, \sigma_H^2)$ e $Y_D \sim N(\mu_D, \sigma_D^2)$, con $\mu_H = -2.5$, $\sigma_H = 4$, $\mu_D = 2.5$ y $\sigma_D = 4$, construida utilizando la ecuación (1.2). A su vez, calculamos el AUC a partir de la fórmula (1.5), obteniendo un valor de 0.81.

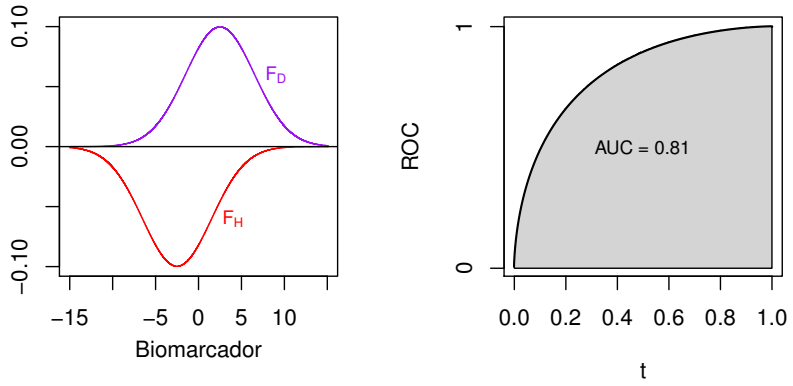


Figura 1.5: Curva ROC y AUC para distribuciones F_D y F_H que siguen el modelo binormal.

Como se puede observar, en este ejemplo el valor de b dado en (1.4) es igual a 1, puesto que las distribuciones del biomarcador en las poblaciones enferma y sana tienen la misma varianza. Si $b = 1$, se tiene que la curva ROC binormal resulta cóncava, lo cual es un atributo deseable ya que asegura que la curva no cruzará la diagonal correspondiente a la función identidad. En cambio, para $b \neq 1$ esta propiedad no se cumple y se pueden producir anomalías donde la curva ROC caiga por debajo de la curva asociada a un test no informativo. Esta situación podría generar cierta preocupación sobre la utilización del modelo binormal ya que, como se menciona en Pepe (2003), es poco probable que ocurra al trabajar con datos y tests de diagnóstico reales. Sin embargo, este tipo de comportamiento indeseable suele producirse en una parte muy pequeña de la curva ROC y en un rango que en muchos casos resulta de poco interés a los fines prácticos. Con lo cual se considera que el modelo binormal podría proporcionar una buena aproximación a las curvas ROC en una gran variedad de casos que ocurren en la práctica y esto ha justificado su extenso uso.

1.5. Estimación de la curva ROC

En esta sección presentaremos algunas propuestas para estimar la curva ROC y sus medidas de resumen. Los conceptos involucrados en los distintos enfoques que abordaremos serán de gran relevancia a lo largo de todo el trabajo, particularmente en el estudio de modelos para estimar las curvas ROC en presencia de covariables.

En adelante, asumiremos que tenemos muestras aleatorias independientes $(Y_{D,i})_{1\leq i\leq n_D}$ e $(Y_{H,i})_{1\leq i\leq n_H}$ que representan los resultados del test en individuos enfermos y sanos respectivamente, cada una idénticamente distribuida de modo que $Y_{D,i} \sim F_D$ e $Y_{H,i} \sim F_H$.

1.5.1. Estimación basada en la distribución empírica

La estimación basada en la distribución empírica consiste en sustituir en la ecuación (1.1) la función de distribución F_D y la función cuantil de F_H por sus estimadores empíricos. En esencia, esta estimación de la curva ROC se basa en el principio de "plug-in" y viene dada por

$$\widehat{\mathrm{ROC}}(t) = 1 - \widehat{F}_D(\widehat{F}_H^{-1}(1-t))\,, \tag{1.6}$$

donde $\widehat{F}_D$ denota la función de distribución empírica asociada a la población enferma y $\widehat{F}_H^{-1}$ denota la función de cuantiles empírica asociada a la población sana.

Recordemos que a partir de las muestras $(Y_{D,i})_{1\leq i\leq n_D}$ e $(Y_{H,i})_{1\leq i\leq n_H}$ se definen las funciones de distribución empíricas en un valor de t como:

$$\begin{aligned} \widehat{F}_H(t) &= \frac{1}{n_H}\sum_{i=1}^{n_H} \mathbf{I}_{\{Y_{H,i}\leq t\}} \\ \widehat{F}_D(t) &= \frac{1}{n_D}\sum_{i=1}^{n_D} \mathbf{I}_{\{Y_{D,i}\leq t\}}\,, \end{aligned} \tag{1.7}$$

donde $\mathbf{I}_A$ denota la función indicadora del conjunto A.

A su vez, en base a estas funciones se pueden obtener los cuantiles estimados empíricamente. En particular, precisamos la función cuantil empírica para la población sana que está definida por:

$$\widehat{F}_H^{-1}(p) = \text{ínf}\{t : \widehat{F}_H(t) \geq p\}\,.$$

Para estimar el área bajo la curva podemos aplicar la definición a la curva ROC estimada empíricamente, es decir:

$$\widehat{\mathrm{AUC}} = \int_0^1 \widehat{\mathrm{ROC}}(t)dt\,.$$

De forma similar se obtienen estimaciones para las otras medidas de resumen.

1.5.2. Modelado paramétrico de las distribuciones del test

Otro enfoque para estimar la curva ROC consiste en modelar de forma paramétrica las distribuciones de Y_D e Y_H, para luego calcular la curva ROC inducida.

Supongamos que asumimos un modelo de distribución paramétrico, con parámetros α y γ para cada distribución del biomarcador en los individuos enfermos y sanos:

$$F_D(y) = F_{\alpha,D}(y), \quad F_H(y) = F_{\gamma,H}(y).$$

Luego, bajo este supuesto podemos estimar α utilizando los datos de la población enferma y γ con los datos de la población sana. Nuevamente basándonos en el formato de la curva ROC deducido en (1.1), la curva ROC estimada resultante es:

$$\widehat{\mathrm{ROC}}_{\widehat{\alpha},\widehat{\gamma}}(t) = 1 - F_{\widehat{\alpha},D}(F^{-1}_{\widehat{\gamma},H}(1-t))\,.$$

A modo de ejemplo, consideremos el modelo binormal que presentamos anteriormente, donde $Y_D \sim N(\mu_D, \sigma_D^2)$ e $Y_H \sim N(\mu_H, \sigma_H^2)$. Estimando, entonces, μ_D, μ_H, σ_D^2 y σ_H^2 se obtienen estimaciones para las funciones de distribución del test en la población enferma y en la sana. Luego, dados estimadores $\widehat{\mu}_D$, $\widehat{\mu}_H$, $\widehat{\sigma}_D$ y $\widehat{\sigma}_H$, tomando las expresiones (1.2) a (1.4) la curva ROC estimada es:

$$\widehat{\mathrm{ROC}}(t) = \phi\left(\frac{\widehat{\mu}_D - \widehat{\mu}_H}{\widehat{\sigma}_D} + \frac{\widehat{\sigma}_H}{\widehat{\sigma}_D}\phi^{-1}(t)\right)\,.$$

Además, el área bajo la curva se puede estimar de la siguiente manera:

$$\mathrm{AUC} = \phi\left(\frac{\widehat{\mu}_D - \widehat{\mu}_H}{\sqrt{\widehat{\sigma}_D^2 + \widehat{\sigma}_H^2}}\right).$$

1.5.3. Modelos de posición y escala semiparamétricos

Los modelos de posición y escala pueden verse como una generalización del modelo anterior, asumiendo que en cada una de las poblaciones el biomarcador viene dado por:

$$Y_D = \mu_D + \sigma_D\epsilon_D \tag{1.8}$$
$$Y_H = \mu_H + \sigma_H\epsilon_H\,, \tag{1.9}$$

donde ϵ_D y ϵ_H son variables aleatorias independientes con media 0, varianza igual a 1 y funciones de distribución acumulada tales que $\epsilon_D \sim G_D$ y $\epsilon_H \sim G_H$. Consideraremos el caso en el que la forma de G_D y G_H no está especificada y por este motivo la estimación resultará semiparamétrica.

Procediendo de modo similar a la deducción de la expresión de la curva ROC (1.2) para el modelo binormal, en el modelo de posición y escala podemos escribir la curva ROC de la siguiente manera:

$$\mathrm{ROC}(t) = 1 - G_D\left(\frac{\mu_H - \mu_D}{\sigma_D} + \frac{\sigma_H}{\sigma_D} G_H^{-1}(1-t)\right). \tag{1.10}$$

Luego, tomando $\widehat{\mu}_D$, $\widehat{\mu}_H$, $\widehat{\sigma}_D$ y $\widehat{\sigma}_H$ estimadores de los parámetros del modelo, estimaremos la curva ROC realizando un plug-in del que resulta:

$$\widehat{\mathrm{ROC}}(t) = 1 - \widehat{G}_D\left(\frac{\widehat{\mu}_H - \widehat{\mu}_D}{\widehat{\sigma}_D} + \frac{\widehat{\sigma}_H}{\widehat{\sigma}_D} \widehat{G}_H^{-1}(1-t)\right),$$

donde $\widehat{G}_D$ y $\widehat{G}_H^{-1}$ corresponden a las estimaciones empíricas de la función de distribución de ϵ_D y de la función cuantil de ϵ_H, basadas en los residuos

$$\left(\frac{Y_{D,i} - \widehat{\mu}_D}{\widehat{\sigma}_D}\right)_{1 \le i \le n_D} \quad \text{y} \quad \left(\frac{Y_{H,i} - \widehat{\mu}_H}{\widehat{\sigma}_H}\right)_{1 \le i \le n_H},$$

respectivamente.

1.5.4. Estimador ROC-GLM

A diferencia de los métodos desarrollados en los apartados anteriores donde se propusieron modelos para las distribuciones de Y_D e Y_H, a continuación presentaremos un enfoque que consiste en modelar la forma de la curva ROC en sí misma.

Para tal fin, necesitaremos introducir la noción de *placement value* (valor de ubicación) que será fundamental en nuestro trabajo y al cual volveremos en los capítulos siguientes. De ahora en adelante notaremos a los "placement values" con las siglas PV. Tomando como referencia la distribución de Y_H, se define el valor de ubicación (PV) para y en relación a la distribución de la población sana como:

$$P(Y_H \ge y) = 1 - F_H(y).$$

Este concepto permite estandarizar mediciones respecto a una distribución de referencia adecuada, en el sentido de que literalmente nos da la ubicación de un valor y en relación a dicha distribución.

A partir de esta definición, la curva ROC se puede interpretar como la función de distribución de los PV del biomarcador en los sujetos enfermos con respecto a la distribución de la población sana. En efecto, si llamamos $\text{PV} = 1 - F_H(Y_D)$:

$$\begin{aligned} P(\text{PV} \leq t) &= P(1 - F_H(Y_D) \leq t) = P(F_H(Y_D) \geq 1 - t) \\ &= P(Y_D \geq F_H^{-1}(1-t)) = 1 - F_D(F_H^{-1}(1-t)) \\ &= \text{ROC}(t). \end{aligned}$$

En la metodología de estimación ROC-GLM se propone cierto modelo para la curva ROC y se suele asumir que adopta una forma paramétrica tal que

$$g(\text{ROC}(t)) = \sum_{s=1}^{S} \alpha_s h_s(t), \tag{1.11}$$

donde la función link g y las S funciones $\{h_1, ..., h_S\}$ son conocidas.

Si consideramos la variable binaria $U_{it} = \mathbf{I}[1 - F_H(Y_{D,i}) \leq t]$, con $\mathbf{I}$ la función indicadora como antes, para $1 \leq i \leq n_D$, tenemos que

$$E(U_{it}) = P(1 - F_H(Y_D) \leq t) = \text{ROC}(t).$$

Luego, el formato (1.11) define un modelo lineal generalizado para la variable U_{it} y el enfoque se basa en ajustar dicho modelo a los datos binarios para estimar los parámetros α.

Para eso, el procedimiento es el siguiente:

- Se elige un conjunto $T \subset (0, 1)$ de valores de t sobre el cual se ajustará el modelo.
- Calculamos los PV empíricos $\{1 - \widehat{F}_H(Y_{D,i}),\ i = 1, ..., n_D\}$, donde $\widehat{F}_H$ es la función de distribución empírica basada en $(Y_{H,i})_{1 \leq i \leq n_H}$.
- Para cada $t \in T$ calculamos las indicadoras binarias basadas en los PV empíricos, $\widehat{PV}_i = 1 - \widehat{F}_H(Y_{D,i})$, de modo que:

  $$\widehat{U}_{it} = \mathbf{I}[\widehat{PV}_i \leq t],$$

  para $i = 1, ..., n_D$.
- Se estiman los parámetros $\{\alpha_1, ..., \alpha_S\}$ utilizando métodos de regresión binaria, tomando la función link g y las covariables definidas por $\{h_1(t), ..., h_S(t)\}$, a partir de los datos:

  $$\left\{ \left(\widehat{U}_{it}, h_1(t), ..., h_S(t) \right),\ t \in T,\ i = 1, ...n_D \right\}$$

Como mencionamos inicialmente, el procedimiento ROC-GLM se basa en los rangos de los datos y solamente asume un modelo para la curva ROC, no requiere modelar las distribuciones del test. Esto representa una ventaja ya que se evitan supuestos que son innecesarios, dado que la curva ROC se ocupa de la relación entre las distribuciones de Y_D e Y_H, y no de las distribuciones en sí mismas.

1.5.5. Robustez de la estimación de la curva ROC

Un aspecto interesante para analizar en relación a la estimación de la curva ROC es su robustez, en el sentido de cuán resistente se presenta ante desviaciones del modelo supuesto, a la vez que produce resultados confiables cuando el modelo se sostiene.

Nos referiremos a Gonçalves *et al.* (2014) para una revisión de la literatura sobre la robustez de la estimación de la curva ROC bajo el modelo binormal. Tal como se menciona allí, distintos autores han argumentado a favor de la robustez del estimador binormal. Sin embargo, Walsh (1997) discute el sentido de esa aceptación y estudia la capacidad del estimador binormal para producir inferencias válidas en circunstancias en las cuales los datos no satisfacen el supuesto de normalidad. A partir de un estudio de simulación, analiza el impacto de tomar datos provenientes de un modelo bilogístico, es decir cuando F_H y F_D corresponden a la distribución logística, en combinación con el estimador binormal, concluyendo que dicho estimador resulta sensible a la especificación errónea del modelo.

A su vez, Greco y Ventura (2011) abordaron el problema de la robustez asumiendo un modelo con funciones de distribución conocidas y desarrollaron estimadores robustos para el área bajo la curva ROC.

En su trabajo de tesis de licenciatura Murrone (2019) muestra un experimento numérico en el cual se estudia la sensibilidad de los estimadores empíricos ante la presencia de una proporción pequeña de datos atípicos. En el mismo, se contaminan las muestras de los sujetos enfermos y sanos provenientes de una distribución normal y se observa que los datos atípicos tienen un impacto en la estimación de las curvas ROC, puesto que dichas estimaciones se desvían de forma considerable de la curva ROC verdadera.

En este sentido, en nuestro trabajo nos interesará investigar la sensibilidad de la estimación de las curvas ROC ante un porcentaje pequeño de datos atípicos cuando disponemos de covariables, enfocándonos en la metodología directa. Además, exploraremos distintas herramientas de estimación robusta con el objetivo de obtener estimadores resistentes a contaminaciones en el conjunto de datos.

Capítulo 2

Curvas ROC con covariables: Método directo

2.1. Introducción

En el capítulo anterior vimos que la curva ROC es un instrumento ampliamente utilizado para evaluar la capacidad discriminatoria de un biomarcador que clasifica entre dos poblaciones o estados.

En muchas situaciones, además del resultado del test, se dispone de variables que contienen información adicional de cada individuo. Como se señala en Pardo-Fernández *et al.* (2014), dichas variables pueden afectar la efectividad y la capacidad discriminatoria del biomarcador, con lo cual es recomendable incorporarlas en el análisis de la curva ROC. Por ejemplo, dentro del campo de la medicina, las caracterísiticas de un paciente, tales como la edad o el sexo, pueden impactar en el resultado de un test y ser de relevancia al momento de determinar un diagnóstico.

Pardo-Fernández *et al.* (2014) detallan dos situaciones diferentes en las cuales resulta importante tener en cuenta la información que proveen las covariables:

- En el primer caso, se considera determinada covariable X (por ejemplo, se puede pensar en el sexo de un paciente) que afecta al resultado del test Y, pero no a su capacidad discriminatoria. Es decir, la separación entre las distribuciones condicionales del resultado del test en ambas poblaciones, enferma y sana, se mantiene igual, independientemente de los valores que tome la covariable X. En este sentido, las curvas ROC condicionadas a la covariable son equivalentes entre sí y, en consecuencia, también lo es la capacidad discriminatoria del test. Sin embargo, al ignorar la distinción que provee la

covariable y considerar todos los datos juntos, se puede derivar en curvas ROC por debajo o por arriba de la curva ROC condicionada. A su vez, las covariables influyen en la elección del valor de umbral c, puesto que para obtener una cierta fracción de verdaderos positivos se requiere de umbrales diferentes.

- Una segunda situación que describen es aquella en la que la capacidad discriminatoria del biomarcador sí se ve afectada al tener en cuenta los valores de la covariable X, ya que la separación entre las distribuciones condicionales del test en las poblaciones sana y enferma cambia según el valor considerado. A la vez, en el ejemplo que dan los autores, cuando se consideran todos los datos juntos se obtiene una curva ROC atenuada respecto a las curvas ROC específicas que muestran una precisión alta de diagnóstico.

En dicho trabajo se pueden ver ilustraciones de cada una de las situaciones mencionadas.

A partir de estos casos, se ve reflejada la necesidad de incorporar la información que proporcionan las covariables en el análisis de las curvas ROC. Como mencionan los autores citados, tenerlas en consideración puede ayudar a identificar cuáles son las poblaciones o las condiciones óptimas en las cuales se debería aplicar un test. El hecho de agrupar los datos y utilizar un valor de umbral en común puede llevar a conclusiones erróneas respecto a la capacidad de clasificación del biomarcador, comparando con la potencial capacidad de diagnóstico que se obtiene cuando se tienen en cuenta los valores de las covariables.

Pepe (2003) también muestra diferentes ejemplos de aplicación en los que las covariables pueden impactar en el resultado de un test o biomarcador. Uno de ellos se enmarca el campo de la radiología, donde resulta de interés considerar como covariable al lector/radiólogo para la interpretación de imágenes. En este caso, suele ser más relevante la curva ROC específica ya que en la práctica los radiólogos utilizan sus propias escalas de calificación en la lectura de imágenes. En general, la curva ROC que combina datos de diferentes lectores no resulta representativa ni refleja correctamente los resultados de lectura de ningún radiólogo en particular.

Teniendo todo esto en consideración, en este capítulo abordaremos la estimación de curvas ROC en presencia de covariables e introduciremos la denominada metodología directa para la incorporación de dicha información adicional, en la que enmarcaremos el resto del trabajo.

2.2. Notación y definiciones generales

En lo que sigue, asumiremos que además del resultado del biomarcador Y, se dispone de información sobre ciertas características de los individuos de ambas poblaciones que viene dada por un vector de covariables $\boldsymbol{X}$. Notaremos con Y_D y $\boldsymbol{X}_D$ al biomarcador y las covariables en la población enferma, respectivamente, y con Y_H y $\boldsymbol{X}_H$ a las correspondientes en la población sana. En este contexto, resultará de interés analizar la capacidad discriminatoria del biomarcador Y en función de los valores que tome el vector de covariables $\boldsymbol{X}$.

Sean $F_{D\boldsymbol{X}}$ y $F_{H\boldsymbol{X}}$ las funciones de distribución de Y_D e Y_H condicionadas a $\boldsymbol{X}$, respectivamente, es decir:

$$\begin{aligned} F_{D\boldsymbol{X}}(y) &= P(Y_D \leq y|\boldsymbol{X}) \\ F_{H\boldsymbol{X}}(y) &= P(Y_H \leq y|\boldsymbol{X}), \end{aligned}$$

donde en $F_{j\boldsymbol{X}}$ se considera la probabilidad condicional dado que $\boldsymbol{X}_j = \boldsymbol{X}$, para $j = D, H$.

De forma análoga a la definición dada para la curva ROC en el Capítulo 1, la curva ROC condicional dado $\boldsymbol{X}$ se define como:

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = 1 - F_{D\boldsymbol{X}}(F^{-1}_{H\boldsymbol{X}}(1-t)), \quad 0 \leq t \leq 1\,. \tag{2.1}$$

De este modo, para los distintos valores de las covariables se pueden obtener diferentes curvas ROC. En el caso de que se tenga una única covariable (esto es, el vector de covariables tiene dimensión 1), cada valor de X en la intersección de los soportes de X_D y X_H tiene asociada una curva $\mathrm{ROC}_X(t)$, lo que define una superficie en función a X y t. En la Figura 2.1 se muestra, a modo de ejemplo, el gráfico de una superficie ROC.

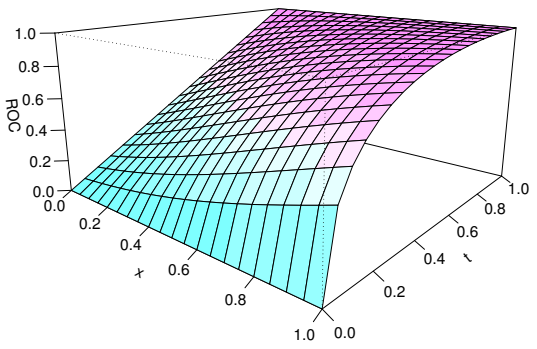


Figura 2.1: Superficie ROC para X univariada.

Al igual que en el caso de curvas ROC sin covariables, existen diversas medidas de resumen que son útiles para caracterizar a las curvas ROC condicionales. Una de las medidas más populares es el *área bajo la curva condicional* ($\text{AUC}_{\boldsymbol{X}}$) que está definida por

$$\text{AUC}_{\boldsymbol{X}} = \int_0^1 \text{ROC}_{\boldsymbol{X}}(t)dt\,.$$

Continuando con el ejemplo en el que se dispone de una sola covariable, al recorrer los valores de X tenemos que AUC_X describe una curva, tal como se muestra en la Figura 2.2. Allí se representan los valores de AUC_X correspondientes a las curvas $\text{ROC}_X(t)$ de la Figura 2.1.

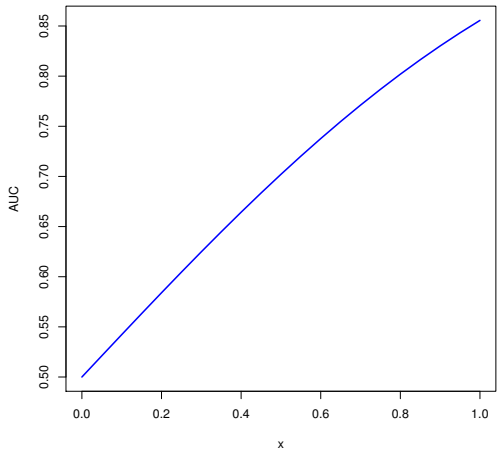


Figura 2.2: Curva AUC_X para X univariada.

2.3. Estimación de la curva ROC condicional

En la sección anterior vimos que se pueden definir curvas ROC condicionadas a valores específicos de las covariables con el fin de tener en cuenta la información adicional que proveen en el análisis. A continuación nos enfocaremos en los procedimientos de estimación de las curvas ROC condicionales.

Existen diversos métodos propuestos para incorporar la información de las covariables a las curvas ROC. Entre ellos, Rodríguez-Álvarez *et al.* (2011) hacen una revisión de dos metodologías en particular dentro del marco general de modelos de regresión, la metodología inducida y la directa. En la metodología inducida la idea central consiste en ajustar modelos de regresión por separado al biomarcador

en cada una de las poblaciones en términos de las covariables, y luego se deriva la curva ROC inducida. En cambio, en la metodología directa a quien se modela es a la curva ROC, es decir, se asume un modelo de regresión lineal generalizado para la curva ROC y los efectos de las covariables en la misma se evalúan directamente.

En el presente trabajo nos enfocaremos en la metodología de estimación directa. Tal como señala Pepe (2003), hay diversas ventajas en modelar de forma directa la curva ROC. Una de las más importantes es que la interpretación de los parámetros del modelo se relaciona directamente con las curvas ROC, ya que esta metodología parametriza los efectos de las covariables sobre la entidad de interés. Otra ventaja es que se pueden evaluar y comparar diferentes biomarcadores en simultáneo, incluso cuando los resultados se miden en distintas unidades o escalas, incorporando al modelo una covariable correspondiente al tipo de test. Además, el modelado directo de la curva ROC permite enfocarse en ciertos rangos de FPF y también habilita la incorporación de interacciones entre las covariables y los valores de FPF.

Para introducir los estimadores de las curvas ROC condicionales, asumiremos que tenemos un vector $\boldsymbol{X}$ de covariables de dimensión p y dos muestras aleatorias independientes $(Y_{D,i}, \boldsymbol{X}_{D,i})_{1\leq i\leq n_D}$ y $(Y_{H,i}, \boldsymbol{X}_{H,i})_{1\leq i\leq n_H}$, cada una idénticamente distribuida, correspondientes a las poblaciones enferma y sana, respectivamente.

2.3.1. Metodología de regresión directa: procedimiento general

Como mencionamos previamente, en la metodología directa se evalúa el efecto de las covariables directamente en la curva ROC y una forma de llevar a cabo esta tarea es a partir de la formulación de modelos dentro del marco del Modelo Lineal Generalizado (GLM). Para describir las caracterísiticas del método y los procedimientos involucrados tomaremos como referencia los trabajos realizados por Pepe (2003) y Rodríguez-Álvarez *et al.* (2011).

Dentro de este contexto, la forma general de la curva ROC viene dada por el modelo de regresión:

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = g\left(\boldsymbol{X}^{\boldsymbol{t}}\boldsymbol{\theta}_0 + h_0(t)\right), \quad t \in (0,1), \tag{2.2}$$

donde g es una función conocida monótona creciente en $(0,1)$ (la inversa de la función link), $\boldsymbol{\theta}_0$ es un vector de parámetros de dimensión p y $h_0(t)$ es una función monótona creciente en $(0,1)$. En caso de que el modelo contenga un término independiente, consideraremos $\boldsymbol{X}^{\boldsymbol{t}} = (1, X_1, ..., X_{p-1})$, siendo X_1, ..., X_{p-1} las componentes del vector de covariables. La ecuación (2.2) define la clase de modelos

que se denominan ROC-GLM.

La función g está especificada como parte del modelo. Algunos ejemplos usuales son la función "probit", con $g \equiv \phi$, siendo ϕ la distribución acumulada de una normal estándar, o la logística donde $g(z) = e^z/(1+e^z)$.

Una de las particularidades que tienen los modelos de regresión ROC-GLM es que la variable dependiente no es directamente observable. Por lo tanto, para ajustar un modelo y estimar la curva ROC es necesario recurrir a otra interpretación, cuya idea central es similar a la desarrollada en la Sección 1.5.4. El enfoque está basado en el concepto clave de *placement value* (valor de ubicación), para el cual ya dimos una definición en la Sección 1.5.4. y la extenderemos ahora al contexto de covariables.

Los "placement values" del biomarcador en la población enferma Y_D en relación a la distribución condicional de la población sana se definen como:

$$\mathrm{PV}_D = 1 - F_{H\boldsymbol{X}}(Y_D).$$

De este modo, análogamente al caso sin covariables, se puede ver que la distribución condicional de los PV_D coincide con la curva ROC, es decir:

$$\begin{aligned} P(\mathrm{PV}_D \leq t|\boldsymbol{X}) &= P(1 - F_{H\boldsymbol{X}}(Y_D) \leq t|\boldsymbol{X}) = P(F_{H\boldsymbol{X}}(Y_D) \geq 1-t|\boldsymbol{X}) \\ &= P(Y_D \geq F^{-1}_{H\boldsymbol{X}}(1-t)|\boldsymbol{X}) = 1 - F_{D\boldsymbol{X}}(F^{-1}_{H\boldsymbol{X}}(1-t)) \\ &= \mathrm{ROC}_{\boldsymbol{X}}(t). \end{aligned}$$

Además, si consideramos la variable binaria $U_{Dt} = \mathbf{I}[\mathrm{PV}_D \leq t]$, donde $\mathbf{I}$ denota la función indicadora, tenemos que

$$E(U_{Dt}|\boldsymbol{X}) = P(\mathrm{PV}_D \leq t|\boldsymbol{X}) = \mathrm{ROC}_{\boldsymbol{X}}(t).$$

En este sentido, la curva ROC condicional puede verse como la esperanza condicional de la variable binaria U_{Dt}. Por lo tanto, el modelo de regresión ROC-GLM definido en (2.2) se puede interpretar como un modelo de regresión para U_{Dt}.

Supongamos que asumimos un formato paramétrico para la función $h_0(t)$, esto es, $h_0(t) = \sum_{s=1}^{S} \alpha_{0s} h_s(t)$, donde $\boldsymbol{\alpha}_0 = (\alpha_{01}, ..., \alpha_{0S})$ es el vector de parámetros desconocidos y las funciones $\{h_1, ..., h_S\}$ son conocidas. Entonces, reemplazando en la ecuación (2.2) se obtiene el siguiente modelo ROC-GLM paramétrico:

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = g\left(\boldsymbol{X}^{\boldsymbol{t}}\boldsymbol{\theta}_0 + \sum_{s=1}^{S} \alpha_{0s} h_s(t)\right), \quad t \in (0,1). \tag{2.3}$$

Basándose en la interpretación de la curva ROC condicional como la esperanza condicional de la variable binaria U_{Dt}, para estimar la curva ROC se puede ajustar el modelo (2.3) a la variable U_{Dt} a partir del siguiente algoritmo:

Paso 1. Elegir un conjunto $T \subset (0,1)$ de valores de t (FPF).

Paso 2. Estimar la distribución condicional $F_{H\boldsymbol{X}}$ sobre la base de la muestra sana $(Y_{H,i}, \boldsymbol{X}_{H,i})_{1\leq i\leq n_H}$. Dicho estimador será notado como $\widehat{F}_{H\boldsymbol{X}}$.

Paso 3. Para cada observación de la población enferma, calcular los PV estimados: $\widehat{PV}_j = 1 - \widehat{F}_{H\boldsymbol{X}_{D,j}}(Y_{D,j})$, $j = 1, ..., n_D$.

Paso 4. Para cada $t \in T$ y cada observación enferma, calcular la indicadora binaria de los PV: $\widehat{U}_{jt} = \mathbf{I}[\widehat{PV}_j \leq t]$, con $t \in T$ y $j = 1, ..., n_D$.

Paso 5. Ajustar el modelo marginal de regresión binaria ROC-GLM (2.3) a los "datos" $\left\{\left(\widehat{U}_{jt}, \boldsymbol{X}_{D,j}, h_1(t), ..., h_S(t)\right),\ t \in T,\ j = 1, ..., n_D\right\}$, obteniendo las estimaciones de los parámetros $\widehat{\boldsymbol{\theta}}$ y $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, ..., \widehat{\alpha}_S)$.

Luego, reemplazando los parámetros de la ecuación (2.3) por los estimadores obtenidos, la curva ROC condicional estimada resultante es

$$\widehat{\mathrm{ROC}}_{\boldsymbol{X}}(t) = g\left(\boldsymbol{X}^t\widehat{\boldsymbol{\theta}} + \sum_{s=1}^{S}\widehat{\alpha}_s h_s(t)\right).$$

Como se puede apreciar, el algoritmo requiere tomar ciertas elecciones y estrategias en las distintas etapas de estimación. A continuación profundizaremos en las técnicas clásicas de estimación involucradas en los Pasos 2 y 5.

2.3.2. Estimación de la distribución $F_{H\boldsymbol{X}}$

Una forma general para incluir las covariables en la estimación de la distribución del biomarcador en la población sana consiste en asumir un modelo de regresión de posición y escala para Y_H, de modo que:

$$Y_H \;=\; \mu_H(\boldsymbol{X}_H) + \sigma_H(\boldsymbol{X}_H)\epsilon_H\,, \tag{2.4}$$

donde, en el enfoque clásico, $\mu_H(\boldsymbol{X}_H) = E(Y_H \mid \boldsymbol{X}_H)$ y $\sigma_H^2(\boldsymbol{X}_H) = \mathrm{Var}(Y_H \mid \boldsymbol{X}_H)$ son las funciones de media y varianza condicionales de la respuesta Y_H dado $\boldsymbol{X}_H$, respectivamente. El error ϵ_H es una variable aleatoria independiente de $\boldsymbol{X}_H$ con media 0 y varianza 1, que está distribuido como $\epsilon_H \sim G_H$.

Vamos a considerar el caso en que las funciones $\mu_H(\boldsymbol{X}_H)$ y $\sigma_H(\boldsymbol{X}_H)$ adoptan una forma paramétrica, mientras que la distribución del error G_H no está especificada. De esta manera, el procedimiento resulta semiparamétrico ya que se combinarán estimadores paramétricos con métodos no paramétricos.

Teniendo en cuenta la independencia entre el error de la regresión y las covariables, podemos escribir a la función de distribución $F_{H\boldsymbol{X}}$ en términos de la distribución del error como

$$F_{H\boldsymbol{X}}(y) = G_H\left(\frac{y-\mu_H(\boldsymbol{X})}{\sigma_H(\boldsymbol{X})}\right), \tag{2.5}$$

puesto que:

$$\begin{aligned} F_{H\boldsymbol{X}}(y) &= P\Big(Y_H \le y \,|\, \boldsymbol{X}\Big) \\ &= P\Big(\mu_H(\boldsymbol{X}_H) + \sigma_H(\boldsymbol{X}_H)\epsilon_H \le y \,|\, \boldsymbol{X}\Big) \\ &= P\left(\epsilon_H \le \frac{y-\mu_H(\boldsymbol{X})}{\sigma_H(\boldsymbol{X})}\right) = G_H\left(\frac{y-\mu_H(\boldsymbol{X})}{\sigma_H(\boldsymbol{X})}\right). \end{aligned}$$

Supongamos que se asume un modelo de regresión lineal homoscedástico para Y_H, es decir, que tomamos $\mu_H(\boldsymbol{X}_H) = \boldsymbol{X}_H^t\boldsymbol{\beta}_H$ y $\sigma_H(\boldsymbol{X}_H) = \sigma_H$, donde $\boldsymbol{\beta}_H$ es un vector de parámetros de dimensión p, de forma que:

$$Y_H = \boldsymbol{X}_H^t\boldsymbol{\beta}_H + \sigma_H\epsilon_H\,.$$

Para el caso en el que el modelo contenga un término independiente, consideraremos $\boldsymbol{X}_H^t = (1, X_{H,1}, ..., X_{H,(p-1)})$ donde $X_{H,1}$, ..., $X_{H,(p-1)}$ corresponden a las componentes del vector de covariables en la población sana.

Entonces, la función de distribución condicional $F_{H\boldsymbol{X}}$ se puede estimar procediendo de la siguiente manera:

- Sobre la base de la muestra de sanos $(Y_{H,i}, \boldsymbol{X}_{H,i})_{1\le i\le n_H}$ estimamos $\boldsymbol{\beta}_H$ a partir del estimador de cuadrados mínimos para el modelo de regresión lineal propuesto, obteniendo así $\widehat{\boldsymbol{\beta}}_H$.
- Estimamos σ_H^2 como:

$$\widehat{\sigma}_H^2 = \frac{\displaystyle\sum_{i=1}^{n_H}\Big(Y_{H,i} - \boldsymbol{X}_{H,i}^t\widehat{\boldsymbol{\beta}}_H\Big)^2}{n_H - p}$$

- Estimamos la función de distribución G_H mediante la distribución empírica sobre la base de los residuos estandarizados de la regresión,

$$\widehat{G}_H(t) = \frac{1}{n_H}\sum_{i=1}^{n_H}\mathbf{I}\left(\frac{Y_{H,i} - \boldsymbol{X}_{H,i}^t\widehat{\boldsymbol{\beta}}_H}{\widehat{\sigma}_H} \le t\right). \tag{2.6}$$

- Por último, considerando la igualdad (2.5) estimamos $F_{H\boldsymbol{X}}(y)$ como

$$\widehat{F}_{H\boldsymbol{X}}(y) = \widehat{G}_H\left(\frac{y - \boldsymbol{X}^t\widehat{\boldsymbol{\beta}}_H}{\widehat{\sigma}_H}\right).$$

Respecto a la función G_H, una alternativa diferente consiste en asumir cierta distribución para los errores y, en ese caso, la metodología resultaría completamente paramétrica. Por ejemplo, podría considerarse $\epsilon_H \sim N(0,1)$ de modo que $G_H = \phi$.

En relación a los estimadores contemplados en los primeros pasos, recordemos que en general, dada una muestra $(Y_1, \boldsymbol{X}_1^t)$, ..., $(Y_n, \boldsymbol{X}_n^t)$, un modelo de regresión lineal viene dado por el formato

$$Y_i = \boldsymbol{X}_i^t\boldsymbol{\beta}_0 + \epsilon_i \qquad 1 \leq i \leq n\,, \tag{2.7}$$

donde Y_i corresponde a la variable de respuesta, $\boldsymbol{X}_i^t = (X_{i1}, ..., X_{ip})$ es el vector de variables independientes o covariables y $\boldsymbol{\beta}_0$ es un vector de parámetros desconocidos de dimensión p. Los ϵ_i son variables aleatorias correspondientes a los errores, independientes de las covariables. En caso de que el modelo contenga un término independiente, se considera $\boldsymbol{X}_i^t = (1, X_{i1}, ..., X_{i(p-1)})$.

De esta manera, el método de cuadrados mínimos para estimar $\boldsymbol{\beta}_0$ consiste en minimizar $\sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})$ respecto a $\boldsymbol{\beta}$, siendo $r_i(\boldsymbol{\beta}) = Y_i - \boldsymbol{X}_i^t\boldsymbol{\beta}$ los residuos del modelo de regresión. Luego, el estimador de cuadrados mínimos de $\boldsymbol{\beta}_0$ es:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})\,. \tag{2.8}$$

Diferenciando respecto de $\boldsymbol{\beta}$ e igualando a cero, se llega a que el estimador $\widehat{\boldsymbol{\beta}}$ verifica

$$\sum_{i=1}^{n} r_i(\widehat{\boldsymbol{\beta}})\boldsymbol{X}_i = \boldsymbol{0}\,. \tag{2.9}$$

Sea $\mathbb{X}$ la matriz de diseño de $n \times p$ compuesta por los elementos X_{ij} y llamemos $\boldsymbol{Y}$ y $\boldsymbol{\epsilon}$ a los vectores cuyos elementos son Y_i y ϵ_i respectivamente, con $1 \leq i \leq n$. Luego, el modelo lineal (2.7) se puede escribir de forma matricial como

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}\,.$$

Las ecuaciones determinadas por la expresión (2.9) son equivalentes a las ecuaciones lineales

$$\mathbb{X}^t\mathbb{X}\boldsymbol{\beta} = \mathbb{X}^t\boldsymbol{Y}\,,$$

que son conocidas como "ecuaciones normales". Si $\mathbb{X}^t\mathbb{X}$ es no singular, la solución es única y viene dada por

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\boldsymbol{Y}\,.$$

Bajo los supuestos de que $\mathbb{X}$ tiene rango completo y los errores ϵ_i verifican

$$E(\epsilon_i) = 0\,, \quad Var(\epsilon_i) = \sigma^2 \quad \text{y} \quad cov(\epsilon_i, \epsilon_j) = 0 \quad i \neq j\,,$$

se cumple que los estimadores $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^t\widehat{\boldsymbol{\beta}}\right)^2$ son insesgados para $\boldsymbol{\beta}_0$ y σ^2 respectivamente, es decir, $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$ y $E(\widehat{\sigma}^2) = \sigma^2$.

Se pueden encontrar más detalles sobre este tema en Seber y Lee (2003).

2.3.3. Modelo Lineal Generalizado: Regresión Binaria

En esta sección haremos una breve revisión del método de regresión binaria que se implementa en el Paso 5 del algoritmo para la estimación de la curva ROC condicional dentro de la metodología directa (Sección 2.3.1).

Dentro de este contexto, llamaremos Y a una variable binaria que toma valores $\{0, 1\}$ y supondremos que $\boldsymbol{X}$ es un vector de covariables de dimensión p. Para modelar el efecto de las covariables sobre Y, podemos asumir que las mismas se relacionan de forma que

$$P(Y = 1|\boldsymbol{X}) = g(\boldsymbol{X}^t\boldsymbol{\theta}_0)\,, \tag{2.10}$$

donde $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ es un vector de parámetros desconocidos y g es una función continua y biyectiva que toma valores en $[0, 1]$. En caso de que el modelo contenga un término "intercept", la primera coordenada de $\boldsymbol{X}$ se toma como 1.

La función g^{-1} se denomina *función de enlace* o *link*, puesto que establece la relación entre ambas componentes del modelo. Como ya hemos mencionado, las funciones de enlace más populares son las correspondientes a la función logística $g(t) = e^t/(1 + e^t)$ y a la función de distribución acumulada normal, denominada "probit", $g(t) = \phi(t)$.

Supongamos que tenemos una muestra $(Y_i, \boldsymbol{X}_i^t)_{1\leq i\leq n}$ que verifica el modelo (2.10). Para simplificar la escritura, dado $\boldsymbol{\theta}$ notaremos

$$p(\boldsymbol{X}_i, \boldsymbol{\theta}) = g(\boldsymbol{X}_i^t\boldsymbol{\theta})\,, \qquad 1 \leq i \leq n.$$

Luego, $p(\boldsymbol{X}_i, \boldsymbol{\theta})$ representa la probabilidad condicional de que Y_i tome valor 1, mientras que $1 - p(\boldsymbol{X}_i, \boldsymbol{\theta})$ es la probabilidad correspondiente a que valga 0. Por lo

tanto, la función de verosimilitud resulta:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{X}_i, \boldsymbol{\theta})^{Y_i} (1 - p(\boldsymbol{X}_i, \boldsymbol{\theta}))^{1-Y_i} .$$

Una forma de estimar $\boldsymbol{\theta}_0$ es a través del método de máxima verosimilitud que consiste en maximizar la función de verosimilitud respecto de $\boldsymbol{\theta}$. Esto último es equivalente a maximizar la función de log-verosimilitud que viene dada por

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} Y_i \log(p(\boldsymbol{X}_i, \boldsymbol{\theta})) + (1 - Y_i) \log(1 - p(\boldsymbol{X}_i, \boldsymbol{\theta})) . \qquad (2.11)$$

De esta manera, el estimador de máxima verosimilitud de $\boldsymbol{\theta}_0$ es

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\max}\, L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\max}\, \ell(\boldsymbol{\theta}) .$$

Al diferenciar la expresión (2.11) respecto de $\boldsymbol{\theta}$ e igualar a cero, se obtiene que el estimador $\widehat{\boldsymbol{\theta}}$ cumple las ecuaciones:

$$\sum_{i=1}^{n} \frac{Y_i - p(\boldsymbol{X}_i, \boldsymbol{\theta})}{p(\boldsymbol{X}_i, \boldsymbol{\theta})\,(1 - p(\boldsymbol{X}_i, \boldsymbol{\theta}))}\, g'(\boldsymbol{X}_i^t \boldsymbol{\theta}) \boldsymbol{X}_i = \mathbf{0} .$$

Típicamente, estas ecuaciones se resuelven numéricamente a partir de métodos tales como el de "Newton-Raphson" o el algoritmo "Fisher-scoring".

Para profundizar en esta temática, ver McCullagh y Nelder (1989).

Capítulo 3

Estimación robusta

Es sabido en la literatura que los estimadores clásicos involucrados en el algoritmo descripto para la metodología de estimación directa de curvas ROC condicionales pueden verse afectados cuando se presentan pequeñas desviaciones del modelo asumido y, en consecuencia, esto podría tener un impacto relevante en la estimación de la curva $\text{ROC}_{\boldsymbol{X}}$.

En el Capítulo 1 mencionamos algunas referencias de autores que estudiaron la sensibilidad de las curvas ROC y desarrollaron estimadores robustos para el caso en el cual no se dispone de covariables. Dentro del contexto de curvas ROC con covariables, en Rodríguez-Álvarez *et al.* (2011) se discute el impacto que tiene el modelado erróneo del efecto de las covariables en la estimación de las curvas ROC para diversos modelos. Entre ellos, analizan el método directo y se evalúa su estabilidad ante la especificación incorrecta del modelo de regresión utilizado para la curva ROC. En un trabajo más reciente, Bianco, Boente y González-Manteiga (2020) estudian la sensibilidad de la estimación de la curva ROC condicional en el marco de la metodología inducida y proponen un método para obtener estimadores robustos que resulten resistentes ante un porcentaje de datos atípicos en las muestras.

Siguiendo en esa línea, a continuación presentaremos diferentes herramientas de estimación robusta con el objetivo de desarrollar un procedimiento confiable y resistente para estimar las curvas $\text{ROC}_{\boldsymbol{X}}$ en el marco del método directo. Nuestro interés está enfocado en obtener métodos robustos que se mantengan estables ante la presencia de un pequeño porcentaje de datos atípicos y, a la vez, resulten eficientes cuando el modelo asumido se sostiene.

3.1. Motivación

Vamos a comenzar analizando algunas situaciones que ponen de manifiesto la influencia que pueden tener los datos atípicos en algunos estimadores clásicos, por más pequeña que sea la proporción en la muestra, obstaculizando la capacidad que tienen para dar una buena aproximación sobre el conjunto de los datos. Los ejemplos que estudiaremos fueron extraídos de Maronna *et al.* (2019).

Primero, consideraremos una situación bien sencilla en la que se dispone de un conjunto de 24 observaciones de cierta variable (Analytical Methods Committee, 1989):

$$\begin{gathered} 2.20,\ 2.20,\ 2.40,\ 2.40,\ 2.50,\ 2.70,\ 2.80,\ 2.90, \\ 3.03,\ 3.03,\ 3.10,\ 3.37,\ 3.40,\ 3.40,\ 3.40,\ 3.50, \\ 3.60,\ 3.70,\ 3.70,\ 3.70,\ 3.70,\ 3.77,\ 5.28,\ 28.95\,. \end{gathered} \tag{3.1}$$

Mirando los datos, salta a la vista que el valor 28.95 se encuentra considerablemente alejado del resto de los valores y, por lo tanto, puede ser tomado como un dato atípico.

Recordemos que, dado un conjunto de datos $\mathbf{x} = (x_1, ..., x_n)$, la media muestral $\bar{x}$ se define como el promedio aritmético

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

y usualmente se utiliza para representar o dar una estimación del valor "central" de los datos. Si calculamos la media para el conjunto definido en (3.1), el resultado es $\bar{x} = 4.28$. Evidentemente, este valor no se encuentra entre la mayor parte de los datos y, en particular, es más grande que la mayoría de ellos (salvo los últimos dos). Eliminando la observación atípica 28.95 la media es $\bar{x} = 3.21$, un valor que parece ser mucho más representativo de la totalidad del conjunto de datos original. Con lo cual, vemos que un solo dato anómalo puede tener una alta influencia en la estimación deseada.

Otra medida muy conocida que podemos utilizar para dar una descripción de los datos es la mediana muestral. Dado el conjunto $\mathbf{x}$, la mediana se define como:

$$\mathrm{Med}(\mathbf{x}) = \begin{cases} x_{(m)} & \text{si } n = 2m-1 \\ \\ \frac{x_{(m)}+x_{(m+1)}}{2} & \text{si } n = 2m \end{cases},$$

donde $(x_{(1)}, ..., x_{(n)})$ corresponden a los datos ordenados. Para el conjunto (3.1) el valor de la mediana muestral con todos los datos completos es 3.38, mientras

que el resultado que se obtiene al eliminar el valor 28.95 es 3.37. Estos resultados son muy parecidos entre sí y, a su vez, se encuentran cercanos al valor de la media cuando no se contempla el dato atípico, lo que nos permite concluir que la mediana no se ve muy afectada por la presencia de dicha observación.

Ahora veamos otro ejemplo que está más vinculado a nuestro objeto de estudio, donde se muestra la sensibilidad del estimador de cuadrados mínimos en un modelo de regresión lineal.

Maronna *et al.* (2019) describen un experimento sobre la velocidad de aprendizaje de las ratas (Bond, 1979) en el cual se registraron los tiempos que tarda una rata en pasar por una caja de lanzadera. Si el tiempo superaba los 5 segundos, se le aplicaba una descarga eléctrica durante el siguiente intento. La Figura 3.1[1] muestra los datos correspondientes al número de descargas recibidas y al tiempo promedio de todos los intentos entre descargas. Como se puede apreciar, salvo por las observaciones 1, 2 y 4, las variables parecen respetar cierta relación lineal. En la Figura 3.1 también están representadas las rectas obtenidas a partir del ajuste lineal por el método de cuadrados mínimos (descripto en la Sección 2.3.2) con el total de los datos, en línea roja, y sin la utilización de los puntos 1, 2 y 4, en línea azul. Podemos ver que la línea roja no provee una buena representación para la mayoría de los datos, dado que la estimación se ve perturbada por equiparar los puntos atípicos con el resto. En cambio, la línea azul parece dar una mejor aproximación general, excepto para los puntos 1, 2 y 4.

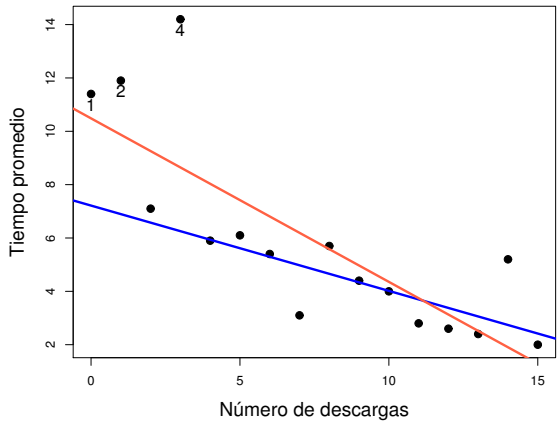


Figura 3.1: Ejemplo: experimento sobre aprendizaje de ratas. Datos y ajuste lineal por el método de cuadrados mínimos con la muestra completa (línea roja) y sin los puntos 1, 2 y 4 (línea azul).

[1]La figura fue obtenida a partir del código **shock.R** del paquete **RobStatTM** de R.

Teniendo en mente las situaciones anteriores, se abre un camino posible para el tratamiento de las observaciones atípicas que consiste en la detección y eliminación de las mismas del conjunto de datos. Sin embargo, como señalan Maronna *et al.* (2019), esta estrategia puede presentar varios problemas, principalmente porque la eliminación de una observación requiere tomar decisiones subjetivas. En general, no suele ser tan evidente cuándo un dato es "suficientemente anómalo" y con la eliminación se corre el riesgo de estar subestimando observaciones que pueden resultar significativas en el análisis.

Por otra parte, en el primer ejemplo vimos que la mediana provee una alternativa robusta para representar el valor central de un conjunto de datos, de forma que se pueden buscar otras posibilidades para la estimación deseada sin la necesidad de eliminar datos de la muestra. Aún así, en ese caso en particular, se puede probar que el rendimiento estadístico de la mediana es peor que el de la media muestral cuando los datos no contienen observaciones atípicas bajo cierto modelo asumido.

Los ejemplos expuestos nos sirven como una pequeña muestra de la sensibilidad que pueden tener los estimadores clásicos ante la presencia de observaciones anómalas. Es por este motivo que en la literatura se han desarrollado diversas alternativas de estimación robustas y resulta de interés la búsqueda de procedimientos que se muestren resistentes cuando la muestra contiene datos atípicos, a la vez que preserven ciertas propiedades estadísticas de los estimadores clásicos bajo el modelo asumido.

3.2. Dispersión y escala

En esta sección presentaremos algunas herramientas de estimación robusta que serán de utilidad posteriormente en el desarrollo de estimadores para modelos de regresión.

3.2.1. Estimadores de dispersión

Una forma usual de medir la variabilidad de un conjunto de datos $\mathbf{x}$ es a través del *desvío estándar muestral* que se calcula como

$$\mathrm{SD}(\mathbf{x}) = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}}.$$

Al igual que con la media muestral, se puede ver que el desvío estándar resulta sensible a la presencia de datos atípicos. Teniendo esto en cuenta, Maronna *et al.* (2019) revisan otros estimadores que se han propuesto en la literatura como alternativas.

Uno de ellos es el denominado *desvío medio absoluto* definido por

$$\mathrm{MD}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

que, aunque es un poco más resistente que el desvío estándar, también se ve influenciado cuando se tienen observaciones atípicas al estar basado en promedios.

A partir de la medida anterior, una alternativa robusta consiste en tomar la mediana en lugar de los promedios, conduciendo a lo que se denomina el estimador MAD (*mediana del desvío absoluto respecto a la mediana*):

$$\mathrm{MAD}(\mathbf{x}) = \mathrm{Med}(|\mathbf{x} - \mathrm{Med}(\mathbf{x})|),$$

donde $|\mathbf{x} - \mathrm{Med}(\mathbf{x})| = (|x_1 - \mathrm{Med}(\mathbf{x})|, ..., |x_n - \mathrm{Med}(\mathbf{x})|)$.

Supongamos que X es una variable aleatoria y que $\mathbf{x}$ corresponde a una realización de la muestra aleatoria $\boldsymbol{X} = (X_1, ..., X_n)$, donde $X_i \sim X$ para todo $1 \leq i \leq n$. De forma análoga a las medidas introducidas para $\mathbf{x}$, se pueden definir el desvío medio absoluto y el correspondiente a la mediana absoluta para una variable aleatoria X a partir de:

$$\mathrm{MD}(X) = E\left(|X - E(X)|\right) \quad \text{y} \quad \mathrm{MAD}(X) = \mathrm{Med}(|X - \mathrm{Med}(X)|).$$

Si $X \sim N(\mu, \sigma^2)$ se cumple que $\mathrm{MAD}(X) = k\sigma$, con $k \approx 0.675$. Luego, para contar con un estimador equiparable al desvío estándar para la distribución normal, se puede normalizar el MAD dividiéndolo por la constante k. Así, dado un conjunto de datos $\mathbf{x}$ se define el MADN como:

$$\mathrm{MADN}(\mathbf{x}) = \frac{\mathrm{MAD}(\mathbf{x})}{0.675}.$$

Otros estimadores de dispersión muy conocidos que están basados en los estadísticos de orden $(x_{(1)}, ..., x_{(n)})$ son el rango y el rango intercuartil. El rango se calcula como $x_{(n)} - x_{(1)}$ y es muy sensible a los datos atípicos, mientras que el rango intercuartil es más estable y viene dado por $\mathrm{IQR}(\mathbf{x}) = x_{(n-m+1)} - x_{(m)}$, con $m = [n/4]$, de manera que IQR corresponde al rango del 50 % central de los datos. Al igual que con el MAD, también es posible normalizar el IQR dividiéndolo por una constante.

3.2.2. M-estimadores de escala

Consideremos ahora un caso particular en el cual las observaciones x_i provienen de una muestra aleatoria $\boldsymbol{X} = (X_1, ..., X_n)$, con $X_i \sim X$, que satisface el modelo

$$X_i = \sigma_0 U_i \,, \tag{3.2}$$

donde las U_i son independientes e idénticamente distribuidas con densidad f_0 y $\sigma_0 > 0$ es un parámetro desconocido. De esta manera, las X_i pertenecen a la familia de distribuciones de escala cuya densidad es

$$\frac{1}{\sigma_0} f_0 \left(\frac{x}{\sigma_0} \right).$$

El estimador de máxima verosimilitud para la escala σ_0 se calcula como:

$$\widehat{\sigma} = \arg \max_{\sigma} \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0 \left(\frac{x_i}{\sigma} \right) .$$

Tomando logaritmo a la función de verosimilitud y derivando respecto de σ, obtenemos que el estimador $\widehat{\sigma}$ verifica

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{x_i}{\widehat{\sigma}} \right) = 1 \,,$$

donde $\rho(t) = -t f_0'(t)/f_0(t)$.

Por ejemplo, si f_0 corresponde a una N(0,1), basándose en el método de máxima verosimilitud, $\rho(t) = t^2$ y resulta que el estimador $\widehat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2}$.

Luego, de manera más general, para estimar la escala σ_0 consideraremos soluciones que satisfagan una ecuación de la forma

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{X_i}{\widehat{\sigma}} \right) = \delta \,, \tag{3.3}$$

donde ρ es una función con determinadas características. Estos estimadores son denominados *M-estimadores de escala*.

A continuación daremos algunas definiciones que se ven involucradas en el trabajo con este tipo de estimadores y a las cuales nos referiremos más adelante.

Definición 3.2.1. *Diremos que una función dada ρ es una ρ-función si verifica las siguientes condiciones:*

1. *$\rho(x)$ es una función par no decreciente en $|x|$,*
2. *$\rho(0) = 0$,*
3. *$\rho(x)$ es creciente para $x > 0$ tal que $\rho(x) < \rho(\infty)$ y*
4. *si ρ es acotada, se asume que $\rho(\infty) = 1$.*

Definición 3.2.2. *Una función ψ será denominada una ψ-función si es la derivada de una ρ-función. En particular, se tiene que ψ es impar y $\psi \geq 0$ para todo $x \geq 0$.*

Entonces, continuando con la definición de un M-estimador de escala, se consideran soluciones de la ecuación (3.3) donde la función ρ es una ρ-función y δ es una constante positiva.

Para n grande, se puede probar que las soluciones de (3.3) convergen a la solución de

$$E\left[\rho\left(\frac{X}{\sigma}\right)\right] = \delta\,,$$

en caso de que sea única.

En muchas ocasiones se suele utilizar una función ρ que es cuadrática cerca del origen, es decir, tal que $\rho'(0) = 0$ y $\rho''(0) > 0$. En ese caso, la ecuación (3.3) es equivalente a

$$\widehat{\sigma}^2 = \frac{1}{n\delta}\sum_{i=1}^{n} W\left(\frac{X_i}{\widehat{\sigma}}\right) X_i^2\,, \tag{3.4}$$

donde W es una función de peso tal que

$$W(x) = \begin{cases} \rho(x)/x^2 & \text{si } x \neq 0 \\ \\ \rho''(0) & \text{si } x = 0\,. \end{cases}$$

La expresión (3.4) sugiere un algoritmo iterativo para computar un M-estimador de escala:

1. Computar un estimador inicial $\widehat{\sigma}_0$ (por ejemplo, el MADN definido en la sección anterior).
2. Para $k \in \{0, 1, 2, ...\}$: dado $\widehat{\sigma}_k$, computar los pesos $w_{k,i} = W(X_i/\widehat{\sigma}_k)$ para $i = 1, ..., n$ y $\widehat{\sigma}_{k+1}$ como

$$\widehat{\sigma}_{k+1} = \sqrt{\frac{1}{n\delta}\sum_{i=1}^{n} w_{k,i} X_i^2}\,. \tag{3.5}$$

3. Detener el procedimiento cuando $|\widehat{\sigma}_{k+1}/\widehat{\sigma}_k - 1| < \eta$.

Si $W(x)$ es acotada, par, continua y no creciente para $x > 0$, se puede ver que la secuencia converge a la solución deseada (ver Maronna *et al.*, 2019, para una demostración).

3.3. Regresión lineal

Como vimos al inicio de este capítulo, el estimador de cuadrados mínimos resulta muy sensible ante la presencia de datos atípicos. En esta sección desarrollaremos otros procedimientos que se han propuesto para obtener métodos robustos para la estimación de los parámetros de un modelo de regresión lineal.

3.3.1. M-estimadores

Vamos a suponer que tenemos una muestra $(Y_i, \boldsymbol{X}_i^t)_{1\leq i\leq n}$ que verifica el modelo lineal dado por (2.7), donde la matriz de diseño $\mathbb{X}$ es fija y de rango completo y los ϵ_i tienen densidad

$$\frac{1}{\sigma_0} f_0\left(\frac{x}{\sigma_0}\right),$$

con σ_0 el parámetro de escala. Luego, las Y_i son independientes y tienen densidad

$$\frac{1}{\sigma_0} f_0\left(\frac{y - \boldsymbol{X}_i^t\boldsymbol{\beta}_0}{\sigma_0}\right).$$

Asumiendo un valor fijo σ_0, el estimador de máxima verosimilitud se obtiene maximizando respecto de $\boldsymbol{\beta}$ la función de verosimilitud

$$L(\boldsymbol{\beta}) = \frac{1}{\sigma_0^n} \prod_{i=1}^{n} f_0\left(\frac{Y_i - \boldsymbol{X}_i^t\boldsymbol{\beta}}{\sigma_0}\right).$$

Esto es equivalente a maximizar la log-verosimilitud (logaritmo de la verosmilitud) o, lo que es lo mismo, minimizar su opuesto; es decir,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{r_i(\boldsymbol{\beta})}{\sigma_0}\right) + \log(\sigma_0)$$

con $\rho_0 = -\log(f_0)$. De esta manera, diferenciando respecto de $\boldsymbol{\beta}$, el estimador $\widehat{\boldsymbol{\beta}}$ verifica la ecuación

$$\sum_{i=1}^{n} \psi_0\left(\frac{r_i(\widehat{\boldsymbol{\beta}})}{\sigma_0}\right) \boldsymbol{X}_i = \boldsymbol{0},$$

donde $\psi_0 = \rho_0' = -f_0'/f_0$.

Algunos ejemplos conocidos son:

- Cuando f_0 corresponde a una normal estándar, $\widehat{\boldsymbol{\beta}}$ coincide con el estimador de cuadrados mínimos (2.8).
- Si f_0 es la densidad de una exponencial doble, $f_0(x) = \frac{1}{2}e^{-|x|}$, $\widehat{\boldsymbol{\beta}}$ satisface

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} |r_i(\boldsymbol{\beta})|. \tag{3.6}$$

  A este estimador $\widehat{\boldsymbol{\beta}}$ se lo denomina *estimador* L_1 y resulta el equivalente a la mediana en el contexto de regresión lineal. A diferencia del estimador de cuadrados mínimos, no siempre es posible encontrar una expresión explícita para una solución de (3.6) por eso, en general, para computarlo existen diversos algoritmos. Se pueden encontrar más detalles al respecto en Maronna *et al.* (2019).

Continuando con el supuesto de que la escala del error σ_0 es conocida, un *M-estimador de regresión* se define como una solución $\widehat{\boldsymbol{\beta}}$ que minimiza la expresión

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\sigma_0}\right). \tag{3.7}$$

Diferenciando respecto de $\boldsymbol{\beta}$ se obtiene la siguiente ecuación para $\widehat{\boldsymbol{\beta}}$:

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\widehat{\boldsymbol{\beta}})}{\sigma_0}\right) \boldsymbol{X}_i = \mathbf{0}\,, \tag{3.8}$$

donde $\psi = \rho'$.

La idea consiste en buscar soluciones a la ecuación anterior que no necesariamente correspondan al estimador de máxima verosimilitud. Si bien en la práctica la escala no suele ser conocida, resulta de interés considerar esta situación a los fines del desarrollo y el estudio de las propiedades de los estimadores. Cuando la escala σ_0 es desconocida, usualmente se suele computar previamente un estimador robusto de escala $\widehat{\sigma}$ y se lo reemplaza en las ecuaciones anteriores, aunque, como veremos luego, también es posible computarlo en simultáneo.

En relación a la elección de ρ y ψ, se suelen considerar ρ- y ψ-funciones, respectivamente. Entre las diversas alternativas que existen, un tipo de funciones

muy populares son las pertenecientes a la familia de *funciones Huber*, que están representadas gráficamente en la Figura 3.2 y se definen como:

$$\rho_k(x) = \begin{cases} x^2 & \text{si } |x| \leq k \\ \\ 2k|x| - k^2 & \text{si } |x| > k \end{cases} \tag{3.9}$$

con derivada $2\psi_k$ tal que

$$\psi_k(x) = \begin{cases} x & \text{si } |x| \leq k \\ \\ \text{sgn}(x)k & \text{si } |x| > k\,, \end{cases} \tag{3.10}$$

donde $\text{sgn}(x)$ es la función que devuelve el signo de x (y vale cero si $x = 0$) y k es una constante de calibración a elección del usuario. Notar que ρ_k es cuadrática en la zona central de los valores de x, pero crece linealmente en el infinito, de forma que esta propuesta se puede pensar como una combinación entre las funciones asociadas al estimador de cuadrados mínimos y al denominado L_1. El valor de k se suele elegir con el fin de garantizar determinada varianza asintótica del estimador cuando los errores tienen una distribución normal.

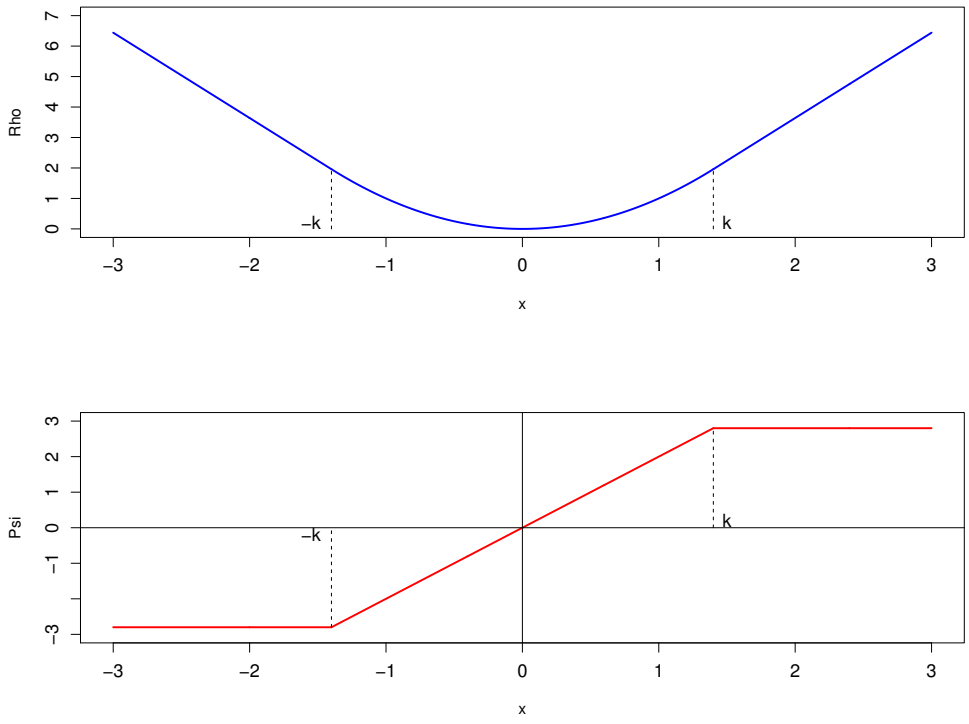


Figura 3.2: Funciones ρ y ψ de Huber.

Otra posibilidad para la elección de la función ψ son las denominadas funciones *redescendientes*, es decir, funciones que decrecen tendiendo a 0 en el infinito. Dentro de ellas, la familia de *funciones bicuadráticas* es una opción ampliamente utilizada, donde

$$\rho_k(x) = \begin{cases} 1 - [1 - (x/k)^2]^3 & \text{si } |x| \leq k \\ \\ 1 & \text{si } |x| > k \end{cases} \tag{3.11}$$

y su derivada es $\rho'_k(x) = 6\psi_k(x)/k^2$ con

$$\psi_k(x) = x \left[1 - \left(\frac{x}{k}\right)^2\right] \mathbf{I}[|x| \leq k]. \tag{3.12}$$

Estas funciones se muestran en la Figura 3.3.

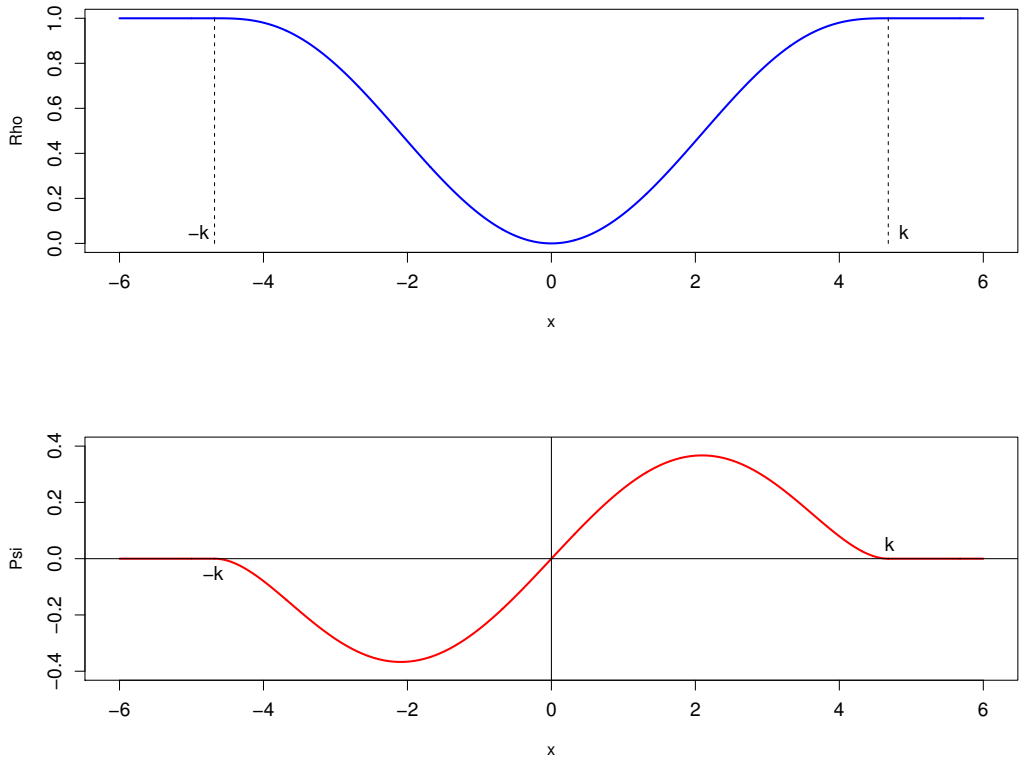


Figura 3.3: Funciones ρ y ψ bicuadráticas.

Para calcular la solución correspondiente a un M-estimador, típicamente se requiere de algún algoritmo iterativo. El esquema general se puede resumir de la siguiente manera:

1. Para empezar, se computa un estimador inicial $\widehat{\boldsymbol{\beta}}_0$ robusto que no dependa de la escala.

2. En relación a la escala, consideraremos los casos en que la misma es:

   i. conocida,

   ii. desconocida y se estima previamente en base a los residuos que provienen de la estimación del Paso 1, o

   iii. desconocida y se computa en simultáneo al estimador $\widehat{\boldsymbol{\beta}}$.

3. Por último, se computa $\widehat{\boldsymbol{\beta}}$ (y $\widehat{\sigma}$ en caso de que se realice en simultáneo) a partir de un procedimiento iterativo.

A continuación profundizaremos en los detalles del algoritmo para cada una de las situaciones planteadas:

• **M-estimadores con escala conocida**

Asumiendo que la escala es conocida y ψ es una función suave, la ecuación (3.8) se puede reescribir introduciendo pesos adecuados como

$$\sum_{i=1}^{n} w_i r_i(\widehat{\boldsymbol{\beta}}) \boldsymbol{X}_i = \sum_{i=1}^{n} w_i \boldsymbol{X}_i (Y_i - \boldsymbol{X}_i^t \widehat{\boldsymbol{\beta}}) = \mathbf{0}\,, \tag{3.13}$$

con $w_i = W(r_i(\widehat{\boldsymbol{\beta}})/\sigma_0)$ y W tal que

$$W(x) = \begin{cases} \psi(x)/x & \text{si } x \neq 0 \\ \\ \psi'(0) & \text{si } x = 0\,. \end{cases}$$

Estas ecuaciones se suelen denominar "ecuaciones normales pesadas" y sugieren un procedimiento iterativo para computar $\widehat{\boldsymbol{\beta}}$ que se detalla a continuación:

1. Computar un estimador inicial $\widehat{\boldsymbol{\beta}}_0$ (por ejemplo, un estimador L_1).

2. Para $k \in \{0, 1, 2, ...\}$, a partir de $\widehat{\boldsymbol{\beta}}_k$ computar $r_{i,k+1} = Y_i - \boldsymbol{X}_i^t \widehat{\boldsymbol{\beta}}_k$ y $w_{i,k+1} = W(r_{i,k+1}/\sigma_0)$, para $i = 1, ..., n$. Luego, computar $\widehat{\boldsymbol{\beta}}_{k+1}$ resolviendo

$$\sum_{i=1}^{n} w_{i,k+1} \boldsymbol{X}_i (Y_i - \boldsymbol{X}_i^t \widehat{\boldsymbol{\beta}}) = \mathbf{0}.$$

3. Detener el procedimiento cuando $\max_i(|r_{i,k} - r_{i,k+1}|)/\sigma_0 < \eta$, donde η es cierto parámetro de tolerancia fijado.

Este procedimiento es llamado "cuadrados mínimos iterativamente ponderados " (o usualmente "IRWLS" por sus siglas en inglés). El algoritmo converge si $W(x)$ es no creciente para $x > 0$.

• M-estimadores con escala estimada preliminarmente

Cuando la escala es desconocida, una posibilidad es computar previamente un estimador de la escala $\widehat{\sigma}$ y resolver la ecuación (3.8) reemplazando σ_0 por $\widehat{\sigma}$.

De esta manera, el algoritmo es similar al descripto anteriormente, con la diferencia de que en el primer paso se computa la estimación de la escala $\widehat{\sigma}$ a partir de los residuos provenientes del estimador inicial $\widehat{\boldsymbol{\beta}}_0$. Maronna *et al.* (2019) proponen, por ejemplo, tomar un estimador inicial L_1 y luego obtener un análogo a la medida MAD normalizada para $\widehat{\sigma}$ a partir de los residuos no nulos de la siguiente forma:

$$\widehat{\sigma} = \frac{1}{0.675}\text{Med}(|r_i(\widehat{\boldsymbol{\beta}}_0)|,\, r_i(\widehat{\boldsymbol{\beta}}_0) \neq 0).$$

• M-estimadores con escala estimada simultáneamente

Otro enfoque posible cuando la escala es desconocida consiste en agregar a la ecuación (3.8) una ecuación simultánea correspondiente a la escala σ_0, obteniendo el sistema:

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\widehat{\boldsymbol{\beta}})}{\widehat{\sigma}}\right)\boldsymbol{X}_i = \mathbf{0}\,, \tag{3.14}$$

$$\frac{1}{n}\sum_{i=1}^{n} \rho_E\left(\frac{r_i(\widehat{\boldsymbol{\beta}})}{\widehat{\sigma}}\right) = \delta\,, \tag{3.15}$$

donde ρ_E es una ρ-función (notada de esa manera para distinguirla de la función ρ de la expresión (3.7)).

El procedimiento para computarlo es similar al caso en que la escala es estimada preliminarmente, excepto que en cada iteración $\widehat{\sigma}$ es actualizado como en (3.5).

Como señalan Maronna *et al.* (2019), cuando ψ es monótona la estimación simultánea resulta menos robusta que la estimación con escala previa. Por otra parte, como las funciones ψ redescendientes tienen la ventaja de disminuir los residuos grandes, recomiendan computar un M-estimador con ρ y ψ pertenecientes a la familia de funciones bicuadráticas combinado con el estimador inicial L_1, sobre el cual se estima la escala previa $\widehat{\sigma}$. Esto implica una alta eficiencia del estimador, concepto que abordaremos a continuación.

Asumamos el modelo lineal (2.7) con la escala σ_0 conocida y los $\epsilon_i \sim \epsilon$ tal que $E\left[\psi\left(\frac{\epsilon}{\sigma_0}\right)\right] = 0$ (que se verifica si la variable ϵ es simétrica y ψ es una ψ-función). Luego, si no hay observaciones con alto "leverage", esto quiere decir que ninguna covariable observada está muy alejada del resto, entonces se verifica que el M-estimador $\widehat{\boldsymbol{\beta}}$ es consistente para $\boldsymbol{\beta}_0$, es decir,

$$\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0.$$

Además, bajo condiciones de regularidad, para n grande se cumple que la distribución de $\widehat{\boldsymbol{\beta}}$ es aproximadamente normal, de forma que

$$\widehat{\boldsymbol{\beta}} \approx N_p\left(\boldsymbol{\beta}_0, \nu(\mathbb{X}^t\mathbb{X})^{-1}\right), \tag{3.16}$$

donde

$$\nu = \sigma_0^2 \frac{E\left[\psi(\epsilon/\sigma_0)^2\right]}{\left(E\left[\psi'(\epsilon/\sigma_0)\right]\right)^2}$$

(para una demostración general, ver Yohai y Maronna (1979)).

Recordemos que si los errores ϵ_i son normales y $\mathbb{X}$ tiene rango completo, entonces el estimador de cuadrados mínimos cumple que

$$\widehat{\boldsymbol{\beta}}_{CM} \sim N_p\left(\boldsymbol{\beta}_0, \sigma_0^2(\mathbb{X}^t\mathbb{X})^{-1}\right).$$

Con lo cual, la matriz de covarianza correspondiente al M-estimador difiere únicamente en una constante en relación a la del estimador de cuadrados mínimos. De esta manera, la *eficiencia* para $\epsilon \sim N(0, \sigma_0^2)$ se define como

$$\text{Eff}(\widehat{\boldsymbol{\beta}}) = \frac{\sigma_0^2}{\nu}.$$

En este sentido, la eficiencia representa la relación entre la varianza asintótica del M-estimador y la correspondiente a uno "óptimo" bajo el supuesto de normalidad, como lo es el estimador de máxima verosimilitud.

Cuando la escala es desconocida y estimada previamente, se puede probar que si $\widehat{\sigma} \xrightarrow{P} \sigma_0$ y se cumple que los errores ϵ_i tienen una distribución simétrica, entonces para n grande la distribución de $\widehat{\boldsymbol{\beta}}$ también se puede aproximar por la expresión (3.16).

3.3.2. S-estimadores

En situaciones en que la matriz de diseño $\mathbb{X}$ no contiene filas $\boldsymbol{X}_i$ con alto leverage y solo las respuestas y_i presentan datos atípicos, un M-estimador con ψ monótona resulta un punto de inicio razonable para computar un estimador robusto

de la escala y un M-estimador redescendiente. Sin embargo, cuando $\mathbb{X}$ es aleatoria los puntos anómalos pueden afectar significativamente a los M-estimadores monótonos.

Por este motivo, a continuación presentaremos una familia de estimadores que no dependen del cómputo de una escala preliminar residual y resultan más robustos que, por ejemplo, el estimador L_1. Estos estimadores, denominados *S-estimadores*, suelen ser un buen punto de partida para computar un M-estimador redescendiente cuando la matriz $\mathbb{X}$ no es determinística.

Los S-estimadores son un caso particular muy importante de estimadores que están basados en una escala robusta para los residuos. Para dar su definición, llamemos $\boldsymbol{r}(\boldsymbol{\beta})$ al vector de residuos, de forma que

$$\boldsymbol{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), ..., r_n(\boldsymbol{\beta}))^t.$$

Supongamos que $\widehat{\sigma}(\boldsymbol{r})$ es un estimador de escala robusto basado en los residuos $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{\beta})$, entonces podemos definir un estimador de regresión en base a $\widehat{\sigma}(\boldsymbol{r})$ de la siguiente manera:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \widehat{\sigma}(\boldsymbol{r}(\boldsymbol{\beta})). \tag{3.17}$$

Se denomina S-estimador a una solución de (3.17) donde $\widehat{\sigma}(\boldsymbol{r})$ es un M-estimador de escala definido en función de $\boldsymbol{r}$ a partir de la ecuación

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right) = \delta\,, \tag{3.18}$$

siendo ρ una ρ-función acotada.

Se puede probar que los S-estimadores con ρ suave son también M-estimadores, donde el estimador de la escala $\widehat{\sigma}$ es computado simultáneamente con $\widehat{\boldsymbol{\beta}}$. Por lo tanto, la distribución asintótica de los S-estimadores es la misma que para los M-estimadores, dada por (3.16).

Desafortunadamente, los S-estimadores con ρ suave que se consideran altamente robustos, en el sentido de que se mantienen estables ante un elevado porcentaje de datos atípicos, no resultan muy eficientes. Sin embargo, son un buen punto de partida para computar los estimadores que introduciremos a continuación, denominados MM-estimadores, que combinan robustez con una alta eficiencia asintótica.

3.3.3. MM-estimadores

Los MM-estimadores son un tipo de M-estimadores que resuelven el problema de minimizar la expresión

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right), \tag{3.19}$$

con ρ acotada y $\widehat{\sigma}$ una escala preliminar robusta. Si ψ es la derivada de ρ, entonces $\widehat{\boldsymbol{\beta}}$ resuelve

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\widehat{\boldsymbol{\beta}})}{\widehat{\sigma}}\right) \boldsymbol{X}_i = \mathbf{0}\,, \tag{3.20}$$

donde ψ es una función redescendiente.

El método general correspondiente a un MM-estimador para aproximar $\widehat{\boldsymbol{\beta}}$ es el siguiente:

1. Se computa un estimador consistente inicial $\widehat{\boldsymbol{\beta}}_0$ que sea robusto (aunque no necesariamente eficiente).

2. Sobre la base de los residuos $r_i\left(\widehat{\boldsymbol{\beta}}_0\right)$, se computa un estimador robusto $\widehat{\sigma}$ de la escala.

3. A partir de $\widehat{\boldsymbol{\beta}}_0$ se utiliza un procedimiento iterativo, como el algoritmo denominado "IRWLS", para hallar una solución a (3.20).

En relación al primer paso, se suele utilizar como estimador inicial un S-estimador con una escala asociada a una función ρ bicuadrática. El estimador de la escala $\widehat{\sigma}$ es un M-estimador de escala que viene dado, en función de $\boldsymbol{r}(\boldsymbol{\beta})$, por

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right) = \delta\,,$$

donde ρ_0 denota la ρ-función acotada correspondiente. Tomando $\delta = 0.5$ se puede probar que el punto de ruptura asintótico del estimador es 0.5. Esto significa, de forma simplificada, que el estimador tolera como máximo un 50 % de datos atípicos en la muestra de forma de seguir brindando alguna información del parámetro a estimar.

Respecto a la elección de la función ρ correspondiente a la expresión (3.19), consideraremos $\rho \leq \rho_0$. Por ejemplo, si ρ^* es la función bicuadrática dada en (3.11) con $k = 1$, podemos tomar

$$\rho_0(r) = \rho^*\left(\frac{r}{c_0}\right) \quad \text{y} \quad \rho(r) = \rho^*\left(\frac{r}{c_1}\right)$$

con $c_1 \geq c_0$ para garantizar que $\rho \leq \rho_0$. El valor de c_0 se elige para asegurar la consistencia del estimador de escala cuando los errores tienen distribución normal, que en el caso de la función bicuadrática es $c_0 = 1.56$. El valor de c_1 se ajusta en función a la eficiencia deseada.

Estos estimadores fueron denominados *MM-estimadores* por Yohai (1987), quien demostró que a partir de este procedimiento $\widehat{\boldsymbol{\beta}}$ resulta consistente. Además, se puede probar que no es necesario encontrar el mínimo absoluto de (3.19) para asegurar una alta eficiencia y un alto punto de rúptura, sino que alcanza con obtener un "buen" mínimo local que sea solución de la ecuación (3.20).

Más detalles pueden encontrarse en Yohai (1987).

3.4. Regresión Binaria

Vamos a asumir que tenemos una muestra $(Y_i, \boldsymbol{X}_i^t)_{1\leq i\leq n}$ que verifica el modelo de regresión binaria dado por (2.10).

Como hemos visto, una forma clásica para estimar el vector de parámetros $\boldsymbol{\theta}_0$ es a través del método de máxima verosimilitud. El estimador de máxima verosimilitud también puede verse como una solución a la minimización de la "deviance", definida como:

$$D(\boldsymbol{\theta}) = \sum_{i=1}^{n} d^2\left(p(\boldsymbol{X}_i, \boldsymbol{\theta}), Y_i\right) , \tag{3.21}$$

donde $d(u, y)$ viene dada por

$$d(u, y) = \{-2[y\log(u) + (1 - y)\log(1 - u)]\}^{1/2}\,\mathrm{sgn}(y - u)\,.$$

La "deviance" representa una medida de discrepancia entre los valores que toma la variable y y su valor esperado.

Con un espíritu similar al de los M-estimadores definidos para modelos de regresión lineal, Pregibon (1981) propone M-estimadores robustos para el modelo de regresión logística (es decir, cuando la función link es la logística), incorporando

una función ρ en la ecuación (3.21). De esta manera, estos estimadores se basan en minimizar la expresión

$$\sum_{i=1}^{n} \rho\left(d^2\left(p(\boldsymbol{X}_i, \boldsymbol{\theta}), Y_i\right)\right), \tag{3.22}$$

donde ρ es una función con un crecimiento más lento que el de la función identidad.

Dentro de este contexto, un M-estimador $\widehat{\boldsymbol{\theta}}$ se puede definir de forma general como una solución a determinada ecuación de formato

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(Y_i, \boldsymbol{X}_i, \boldsymbol{\theta}) = \mathbf{0}.$$

Un M-estimador de este tipo se denomina "consistente Fisher" si se verifica que

$$E[\boldsymbol{\Psi}(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)] = \mathbf{0}.$$

Bianco y Yohai (1996) observaron que los estimadores dados por (3.22) no son consistentes en el sentido de Fisher cuando las $\boldsymbol{X}_i$ son aleatorias. Por este motivo, propusieron agregar un término de corrección a la ecuación, de forma que el estimador $\widehat{\boldsymbol{\theta}}$ se puede obtener minimizando

$$\sum_{i=1}^{n} \left[\rho\left(d^2\left(p(\boldsymbol{X}_i, \boldsymbol{\theta}), Y_i\right)\right) + q(p(\boldsymbol{X}_i, \boldsymbol{\theta}))\right], \tag{3.23}$$

donde ρ es no decreciente y acotada y $q(u) = \nu(u) + \nu(1-u)$, con

$$\nu(u) = 2\int_0^u \psi(-2\log(t))dt$$

y $\psi = \rho'$. Se puede probar que bajo el modelo logístico estos estimadores son consistentes en el sentido de Fisher.

Croux y Haesbroeck (2003) detallan condiciones para la función ρ que garantizan un mínimo finito de la función objetivo dada en (3.23) para muestras con observaciones que tienen "overlapping" (es decir, que no están completamente separadas por un hiperplano) y sugieren tomar ψ dentro de la familia de funciones

$$\psi_c^{CH}(u) = \exp\left(-\sqrt{\max(u, c)}\right).$$

A partir de esta propuesta se pueden obtener estimadores que resultan mucho más robustos que el estimador de máxima verosmilitud.

Cantoni y Ronchetti (2001) desarrollaron otro enfoque que consiste en una robustificación del método de "quasi-verosimilitud" para la estimación de los parámetros de un Modelo Lineal Generalizado (notado "GLM" por sus siglas en inglés).

Dentro del marco de los modelos GLM, se suelen considerar Y_i provenientes de una familia exponencial tales que $E[Y_i] = \mu_i$, $Var[Y_i] = V(\mu_i)$ y

$$\eta_i = g(\mu_i) = \boldsymbol{X}_i^t \boldsymbol{\theta}_0, \quad i = 1, ..., n\,, \tag{3.24}$$

donde $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ es el vector de parámetros, $\boldsymbol{X}_i \in \mathbb{R}^p$ corresponde al vector de covariables y g es la función de enlace.

Cabe destacar que la familia de distribuciones exponenciales contiene a la distribuición Bernoulli que es la que nos interesa en el contexto de regresión binaria.

El estimador de quasi-verosimilitud propuesto por Wedderburn (1974) para el modelo (3.24) se define como una solución al sistema de ecuaciones determinado por

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i}{V(\mu_i)} \mu_i' = \mathbf{0}\,, \tag{3.25}$$

donde $\mu_i(\boldsymbol{\theta}) = g^{-1}(\boldsymbol{X}_i^t \boldsymbol{\theta})$ y $\mu_i' = \frac{\partial}{\partial \beta} \mu_i$.

Sobre la base de este tipo de estimadores, Cantoni y Ronchetti (2001) proponen una clase general de M-estimadores determinados por la ecuación

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(Y_i, \mu_i) = \mathbf{0}\,, \tag{3.26}$$

donde $\boldsymbol{\Psi}(Y_i, \mu_i) = \nu(Y_i, \mu_i) w(\boldsymbol{X}_i) \mu_i' - a(\boldsymbol{\theta})$ y $a(\boldsymbol{\theta})$ viene dado por

$$a(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} E[\nu(Y_i, \mu_i)] w(\boldsymbol{X}_i) \mu_i'\,,$$

con la esperanza tomada respecto de la distribución condicional de $Y|\boldsymbol{X}$.

Con el fin de garantizar la robustez del estimador, se considera una función $\boldsymbol{\Psi}$ acotada para regular la influencia de los datos atípicos. De esta manera, $\nu(\cdot, \cdot)$ es una función acotada que se introduce para controlar las desviaciones en Y, mientras que los puntos de alto leverage en $\boldsymbol{X}$ se tratan con la función de peso $w(\cdot)$. El término de corrección $a(\boldsymbol{\theta})$ se incorpora para asegurar la consistencia Fisher del estimador. Vale la pena destacar que con la elección de $\nu(Y_i, \mu_i) = \frac{Y_i - \mu_i}{V(\mu_i)}$ y $w(\boldsymbol{X}_i) = 1$ se obtiene el estimador de quasi-verosimilitud clásico.

Los autores citados describen el caso particular del estimador dado por (3.26) para el modelo binomial, en el cual consideran

$$\nu(Y_i, \mu_i) = \psi_k(r_i) \frac{1}{V^{1/2}(\mu_i)} ,$$

donde $r_i = \frac{Y_i - \mu_i}{V^{1/2}(\mu_i)}$ (los r_i son denominados "Residuos de Pearson") y ψ_k es la función de Huber definida en (3.10). La constante k se suele elegir de forma de asegurar la eficiencia asintótica deseada.

3.5. Distribución empírica pesada adaptativa

En esta sección abordaremos otro enforque concentrándonos en la etapa de estimación no paramétrica que está involucrada en la metodología directa para estimar curvas ROC con covariables.

Sean Y una variable aleatoria y $\boldsymbol{X}$ un vector de covariables de dimensión p que se relacionan a partir del modelo lineal:

$$Y \quad = \quad \boldsymbol{X}^t \boldsymbol{\beta}_0 + \sigma_0 \epsilon \, , \tag{3.27}$$

donde el error ϵ es una variable aleatoria independiente de $\boldsymbol{X}$ con media 0 y varianza 1. Además, asumiremos que $\epsilon \sim G$, donde G es una función simétrica alrededor del 0.

Como vimos en el Capítulo 2, la función de distribución condicional de Y dado $\boldsymbol{X}$, $F_{\boldsymbol{X}}$, se puede escribir como

$$F_{\boldsymbol{X}}(y) = G\left(\frac{y - \boldsymbol{X}^t \boldsymbol{\beta}_0}{\sigma_0}\right).$$

Basándonos en esta igualdad, si $(Y_i, \boldsymbol{X_i}^t)_{1 \leq i \leq n}$ es una muestra que verifica el modelo (3.27), una forma de estimar la función de distribución condicional $F_{\boldsymbol{X}}$ es a partir de

$$\widehat{F}_{\boldsymbol{X}}(y) = \widehat{G}\left(\frac{y - \boldsymbol{X}^t \widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}\right) ,$$

donde $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}$ son los estimadores clásicos para un modelo de regresión lineal y $\widehat{G}$ se obtiene mediante la distribución empírica sobre los residuos estandarizados de la regresión.

Ya hemos observado previamente que los estimadores clásicos de los parámetros de un modelo de regresión lineal se ven afectados cuando la muestra contiene datos

atípicos. Por lo tanto, con el fin de obtener una estimación robusta de $F_{\boldsymbol{X}}$ para empezar será necesario incorporar estimadores robustos de los parámetros $\boldsymbol{\beta}_0$ y σ_0, como los presentados en las secciones anteriores. Sin embargo, esto solo no resulta suficiente ya que la presencia de residuos grandes puede tener una influencia importante en la distribución empírica clásica. Por este motivo, en lo que sigue introduciremos una clase de estimadores empíricos robustos que están basados en la asignación de pesos a los residuos de mayor tamaño, con el objetivo de disminuir su impacto en la función de distribución estimada.

Considerando $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}$ estimadores robustos de los parámetros del modelo de regresión (3.27), los residuos estandarizados se definen como

$$r_i = \frac{Y_i - \boldsymbol{X}_i^t \widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}.$$

A partir de esta definición, podemos notar que valores altos de $|r_i|$ sugieren que el par $(Y_i, \boldsymbol{X}_i^t)$ representa una observación anómala. Si se asume un modelo normal estándar para los errores, resulta razonable considerar atípicos aquellos puntos que tienen $|r_i| \geq c$, donde el valor de corte c típicamente se elige como 2.5. De esta manera, tomando la propuesta sugerida en Bianco, Boente y González-Manteiga (2020), una vez computados los residuos de un ajuste robusto, buscaremos amortiguar aquellos valores que se alejan de la mayoría de los datos asignádoles un peso menor que al resto de las observaciones. Es decir, consideraremos una distribución empírica pesada dada por

$$\widehat{G}(t) = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \mathbf{I}_{\{r_i \leq t\}} \quad \text{con} \quad w_i = w\left(\frac{|r_i|}{c}\right), \tag{3.28}$$

donde $\mathbf{I}$ denota la función indicadora y la función de peso $w : [0, \infty) \to [0, 1]$ es una función no creciente, continua a derecha, continua en un entorno del 0, tal que $w(0) = 1$, $w(u) > 0$ para $0 < u < 1$ y $w(u) = 0$ si $u \geq 1$. Esta última condición asegura que $w_i = 0$ si $|r_i| \geq c$, de forma que las observaciones con residuos grandes serán completamente eliminadas en el proceso de amortiguación. Una de las funciones de peso más comúnmente utilizadas es la función $w(u) = \mathbf{I}\{u < 1\}$.

En relación a la elección del valor de corte c, tal como proponen Bianco, Boente y González-Manteiga (2020), tomaremos los valores adaptativos de corte definidos por Gervini y Yohai (2002) en el marco del desarrollo de estimadores robustos y eficientes para modelos de regresión lineal. Para la construcción de dichos valores, consideraremos la función de distribución empírica sobre el módulo de los residuos estandarizados definida por

$$\widehat{G}_n^+(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}_{\{|r_i| \leq t\}}$$

y $G^+(t)$ la función de distribución de los valores absolutos de los errores. Como sugieren Gervini y Yohai (2002), una forma de detectar observaciones anómalas es a través de la compración entre las funciones $\widehat{G}_n^+(t)$ y $G^+(t)$. De este modo, si para un valor grande de t se cumple que $\widehat{G}_n^+(t) < G^+(t)$, entonces la proporción de residuos que en valor absoluto exceden a t es mayor que la proporción teórica, indicando que la muestra contiene datos atípicos.

Dado que en la práctica la distribución verdadera de los errores es desconocida, se suele asumir una distribución hipotética G_0 como, por ejemplo, la correspondiente a una distribución normal estándar (es decir, $G_0 = \phi$). Luego, Gervini y Yohai (2002) definen como medida para la proporción de datos atípicos en la muestra

$$d_n = \sup_{t \geq \eta} \left\{ G_0^+(t) - \widehat{G}_n^+(t) \right\}^+ ,$$

donde $\{\cdot\}^+ = \text{máx}(\cdot, 0)$ denota la parte positiva, G_0^+ es la distribución de la variable $|V|$ cuando $V \sim G_0$ y η es un cuantil grande de G_0^+, usualmente elegido como $\eta = 2.5$. Si $|r|_{(1)} \leq |r|_{(2)} \leq ... \leq |r|_{(n)}$ son los estadísticos de orden del valor absoluto de los residuos estandarizados, se verifica que

$$d_n = \max_{i > i_0} \left\{ G_0^+(|r|_{(i)}) - \frac{(i-1)}{n} \right\}^+ ,$$

con $i_0 = \text{máx}\{i : |r|_{(i)} < \eta\}$.

Así, se puede tomar como valor de corte

$$c_n = |r|_{(i_n)} = \text{mín} \left\{ t : \widehat{G}_n^+(t) \geq 1 - d_n \right\} ,$$

donde $i_n = n - [nd_n]$ y, de esta manera, se eliminan las $[nd_n]$ observaciones con mayores residuos.

Por último, con este valor de corte adaptativo y a partir de una función de peso w como la que describimos previamente, se definen los pesos

$$w_i = w\left(\frac{|r_i|}{c_n}\right)$$

y se computa la distribución empírica pesada adaptativa dada por (3.28).

3.6. Procedimiento robusto para curvas ROC con covariables

El enfoque de regresión directa para estimar curvas ROC con covariables involucra diversas etapas de estimación que, como vimos, pueden verse seriamente

afectadas cuando la muestra contiene datos atípicos.

A lo largo de este capítulo recorrimos las distintas dificultades que pueden presentarse y desarrollamos alternativas robustas tanto en el marco paramétrico como en el no paramétrico.

Teniendo en consideración todas estas ideas, en esta sección presentaremos una propuesta robusta para estimar las curvas ROC en presencia de covariables sobre la base de la metodología de regresión directa que combina varios de los métodos robustos descriptos previamente.

El procedimiento consiste en una adaptación del algoritmo presentado en la Sección 2.3.1 donde esencialmente se modifican los Pasos 2 y 5, tal como se describe a continuación:

Paso 1R. Se elige un conjunto $T \subset (0,1)$ de valores de t (FPF).

Paso 2R. Estimamos de forma robusta la distribución $F_{H\boldsymbol{X}}$ sobre la base de la muestra sana $(Y_{H,i}, \boldsymbol{X}_{H,i})_{1\leq i\leq n_H}$. Suponiendo el modelo de regresión lineal

$$Y_H = \boldsymbol{X}_H^t \boldsymbol{\beta}_H + \sigma_H \epsilon_H$$

como en la Sección 2.3.2.:

- Computamos estimadores robustos $\widehat{\boldsymbol{\beta}}_H$ y $\widehat{\sigma}_H^2$ para el modelo de regresión lineal a partir de la muestra $(Y_{H,i}, \boldsymbol{X}_{H,i})_{1\leq i\leq n_H}$.
- Obtenemos los residuos estandarizados para el modelo de regresión

$$r_i = \frac{Y_{H,i} - \boldsymbol{X}_{H,i}^t \widehat{\boldsymbol{\beta}}_H}{\widehat{\sigma}_H}.$$

  A partir de estos residuos, estimamos la distribución de los errores G_H de forma robusta, a quien notaremos como $\widehat{G}_H$.
- Estimamos $F_{H\boldsymbol{X}}(y)$ como

$$\widehat{F}_{H\boldsymbol{X}}(y) = \widehat{G}_H\left(\frac{y - \boldsymbol{X}^t \widehat{\boldsymbol{\beta}}_H}{\widehat{\sigma}_H}\right).$$

Paso 3R. Para cada observación de la población enferma, se calculan los PV estimados a partir del estimador robusto computado en el paso anterior:

$$\widehat{PV}_j = 1 - \widehat{F}_{H\boldsymbol{X}_{D,j}}(Y_{D,j}),\ j = 1, ..., n_D.$$

Paso 4R. Para cada $t \in T$ y cada observación enferma, calculamos la indicadora binaria de los PV: $\widehat{U}_{jt} = \mathbf{I}[\widehat{PV}_j \leq t]$, con $t \in T$ y $j = 1, ..., n_D$.

Paso 5R. A partir de los datos $\left\{\left(\widehat{U}_{jt}, \boldsymbol{X}_{D,j}, h_1(t), ..., h_S(t)\right),\ t \in T,\ j = 1, ..., n_D\right\}$, estimamos los parámetros $\boldsymbol{\theta}_0$ y $\boldsymbol{\alpha}_0$ del modelo ROC-GLM (2.3) utilizando métodos robustos para modelos de regresión binaria. Obtenemos de esta manera los estimadores $\widehat{\boldsymbol{\theta}}$ y $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, ..., \widehat{\alpha}_S)$.

Por último, al igual que con los estimadores clásicos, realizamos un plug-in de los estimadores robustos computados en el último paso, obteniendo

$$\widehat{\mathrm{ROC}}_{\boldsymbol{X}}(t) = g\left(\boldsymbol{X}^t\widehat{\boldsymbol{\theta}} + \sum_{s=1}^{S}\widehat{\alpha}_s h_s(t)\right).$$

El objetivo de robustificar los Pasos 2 y 5 se puede llevar a cabo de distintas maneras. Con respecto al Paso 2R, dentro de las opciones robustas que desarrollamos en la Sección 3.3, en nuestro trabajo optamos por implementar un MM-estimador para la estimación de los parámetros del modelo de regresión lineal. A su vez, para estimar G_H utilizamos la distribución empírica pesada adaptativa dada en (3.28). En relación al Paso 5R, consideramos la propuesta robusta de Bianco y Yohai (1996) para modelos de regresión logística y los estimadores robustos definidos en Cantoni y Ronchetti (2001) para modelos GLM. En el siguiente capítulo, donde mostramos un estudio de simulación, daremos más detalles sobre la implementación de estos estimadores.

Capítulo 4

Estudio de simulación

En este capítulo se presentan los resultados de un estudio de simulación realizado para comparar los estimadores de las curvas ROC introducidos en los capítulos anteriores cuando se dispone de covariables.

El objetivo consiste en estudiar la sensibilidad que tiene la metodología de estimación directa clásica ante un porcentaje pequeño de datos atípicos y, a su vez, evaluar el desempeño del procedimiento propuesto comparándolo con los estimadores clásicos en diferentes esquemas de contaminación.

Las simulaciones fueron implementadas en R donde utilizamos las librerías **MASS** y **robustbase**. En todos los casos, realizamos $Nrep = 1000$ replicaciones generando muestras de datos de tamaño $n_H = n_D = n = 100$ y $n_H = n_D = n = 200$. Consideramos el siguiente escenario basándonos en las condiciones presentadas en Carvalho *et al.* (2013):

$$Y_{H,i} = 0.5 + X_{H,i} + \sigma_H \epsilon_{H,i}, \tag{4.1}$$

$$Y_{D,i} = 2 + 4X_{D,i} + \sigma_D \epsilon_{D,i}, \tag{4.2}$$

para todo $1 \leq i \leq n$, donde las covariables $X_{H,i}$ y $X_{D,i}$ fueron generadas con distribución $U(-1, 1)$, los errores $\epsilon_{H,i}$ y $\epsilon_{D,i}$ con distribución $N(0, 1)$ y $\sigma_H = 1.5$ y $\sigma_D = 2$, siendo los errores y las covariables independientes entre sí.

Bajo este escenario la verdadera curva ROC condicional es:

$$\begin{aligned} \mathrm{ROC}_X(t) &= 1 - \phi\left(\frac{-1.5 - 3X}{2} + 0.75\phi^{-1}(1-t)\right) \\ &= \phi\left(\frac{1.5 + 3X}{2} + 0.75\phi^{-1}(t)\right). \end{aligned}$$

Los estimadores fueron calculados sobre una grilla de puntos equiespaciados,

tomando los $\{x_i\}_{i=1,\dots,n_x}$ dentro del intervalo $I = [-1, 1]$ con una distancia de 0.02 entre sí y los $\{t_j\}_{j=1,\dots,n_t}$ entre 0.02 y 0.98 con una distancia de 0.02.

Utilizamos las siguientes medidas para evaluar el desempeño de los estimadores:

i. El Error Cuadrático Medio (notado "MSE" por sus siglas en inglés) que viene dado por $\text{MSE} = \dfrac{1}{n_x}\displaystyle\sum_{i=1}^{n_x}\frac{1}{n_t}\sum_{1\le j\le n_t}\left(\widehat{\text{ROC}}_{x_i}(t_j) - \text{ROC}_{x_i}(t_j)\right)^2$.

ii. La medida $\text{MAX} = \displaystyle\max_{1\le i\le n_x}\max_{1\le j\le n_t}\left|\widehat{\text{ROC}}_{x_i}(t_j) - \text{ROC}_{x_i}(t_j)\right|$.

En relación a los estimadores implementados, en todo el estudio nos basamos en la metodología de regresión directa desarrollada en el Capítulo 2, contemplando el modelo:

$$\text{ROC}_X(t) = g\left(\theta_0 + \theta_1 X + h_0(t)\right).$$

Consideramos el caso en que la función h_0 adopta la forma paramétrica $h_0(t) = \alpha_1 h_1(t)$ con $h_1(t) = \phi^{-1}(t)$, donde ϕ denota la función de distribución acumulada de una variable aleatoria normal estándar.

En todos los casos asumimos un modelo de regresión lineal para la variable de respuesta en la población sana de forma que:

$$Y_{H,i} = \mu_H(X_{H,i}) + \sigma_H \epsilon_{H,i},$$

con $\mu_H(X_{H,i}) = \beta_0 + \beta_1 X_{H,i}$ para todo $1 \le i \le n$. De este modo, la función de distribución condicional correspondiente a los individuos sanos $F_{H,X}$ verifica que

$$F_{H,X}(y) = G_H\left(\frac{y - \mu_H(X)}{\sigma_H}\right), \tag{4.3}$$

donde G_H es la función de distribución de los errores $\epsilon_{H,i}$.

Con respecto a la metodología de estimación directa clásica de las curvas ROC, implementamos los pasos descriptos en el Capítulo 2. Para el modelo de regresión lineal generalizado, consideramos como función link la función "probit", es decir, $g = \phi$, y para estimar los parámetros $\widehat{\theta}_0$, $\widehat{\theta}_1$ y $\widehat{\alpha}_1$ computamos los estimadores de máxima verosimilitud correspondientes. Al momento de estimar la función de distribución acumulada $F_{H,X}$, utilizamos el método de *plug-in* basándonos en la ecuación (4.3). Los estimadores de los parámetros $\widehat{\beta}_0$, $\widehat{\beta}_1$ y $\widehat{\sigma}_H$ que computamos son los estimadores usuales de mínimos cuadrados para modelos de regresión lineal y la función $\widehat{G}_H$ fue estimada mediante la distribución empírica sobre la base de los residuos estandarizados de la regresión. De este modo, consideramos:

$$\widehat{F}_{H,X}(y) = \widehat{G}_H\left(\frac{y - \widehat{\mu}_H(X)}{\widehat{\sigma}_H}\right), \quad \text{con} \quad \widehat{\mu}_H(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

En relación a la propuesta de estimación robusta, seguimos los pasos detallados al final del Capítulo 3. Para estimar la función de distribución $F_{H,X}$ realizamos un procedimiento similar al del método directo clásico, pero reemplazamos los estimadores de la regresión lineal y de la estimación empírica por estimadores robustos. Los estimadores robustos que utilizamos para el modelo de regresión lineal son los MM-estimadores introducidos en Yohai (1987), con la escala computada a partir de una función ρ bicuadrática de constante $k_0 = 1.54$ y un M-estimador redescendiente asociado a una función ψ bicuadrática con $k_1 = 4.68$. Para la estimación de $\widehat{G}_H$ consideramos la distribución empírica pesada adaptativa que presentamos en el Capítulo 3, a partir de la extensión de los estimadores robustos definidos por Gervini y Yohai (2002). Por último, en la parte correspondiente al modelo lineal generalizado contemplamos dos alternativas de estimación robustas, asumiendo distintas funciones link:

i. los estimadores robustos definidos en Cantoni y Ronchetti (2001), computados a partir de una función ψ de Huber con constante $k = 1.345$, para la función link $g = \phi$ (al cual nos referiremos como método "Robusto 1"); y

ii. los estimadores robustos correspondientes a la propuesta de Bianco y Yohai (1996), utilizando la función ψ definida por Croux y Haesbroeck (2003) de constante $c = 0.5$, para la función link logística $g(z) = \frac{e^z}{1+e^z}$ (que llamaremos método "Robusto 2").

Para cada una de las replicaciones obtuvimos los estimadores de las curvas ROC condicionales evaluados en la grilla de puntos definida anteriormente. Se calcularon los índices de calidad MSE y MAX y, luego, efectuamos sus promedios sobre el conjunto de todas las replicaciones. En el Cuadro 4.1 se presentan los resultados del MSE y MAX obtenidos para cada estimador a partir de las muestras de datos originales bajo el modelo definido en (4.1) y (4.2).

	$n = 100$			$n = 200$		
	Clásico	Robusto 1	Robusto 2	Clásico	Robusto 1	Robusto 2
MSE	0.0029	0.0032	0.0033	0.0013	0.0014	0.0015
MAX	0.1300	0.1372	0.1412	0.0868	0.0906	0.0940

Cuadro 4.1: Promedio de MSE y MAX sobre las replicaciones bajo el modelo (4.1) y (4.2).

Por otra parte, como ya hemos mencionado, una de las medidas de resumen más utilizadas para evaluar la capacidad discriminatoria de un biomarcador es el AUC y, en presencia de covariables, resulta de interés analizar el área bajo la

curva condicional notada por AUC_X. Por lo tanto, en cada replicación calculamos el estimador $\widehat{\text{AUC}}_x$ para los distintos valores de x de la grilla y obtuvimos una curva asociada a cada muestra. En la Figura 4.1 se muestran los boxplots funcionales de las $\widehat{\text{AUC}}_x$ correspondientes a cada estimador, junto con la verdadera curva AUC_x graficada en verde.

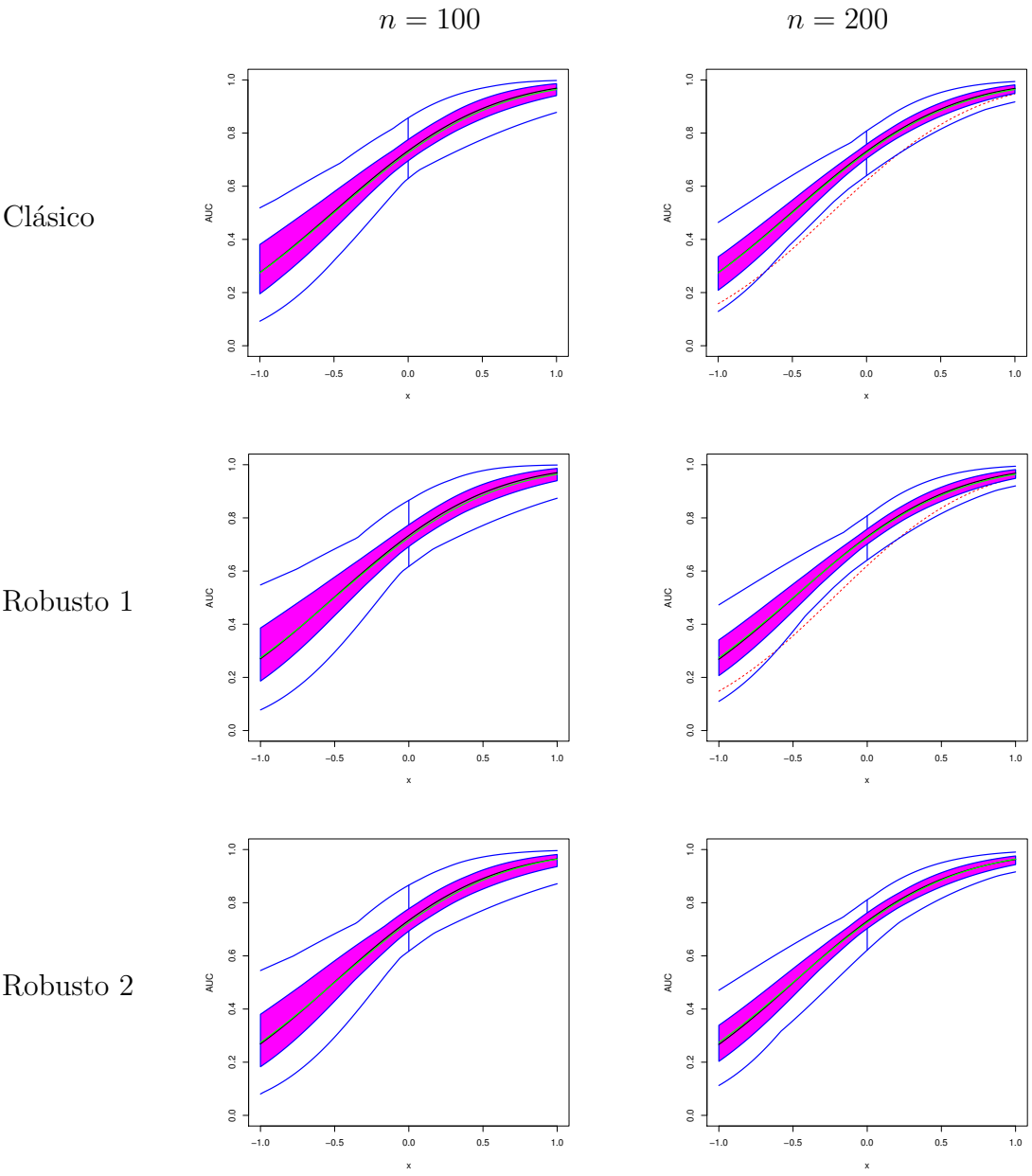
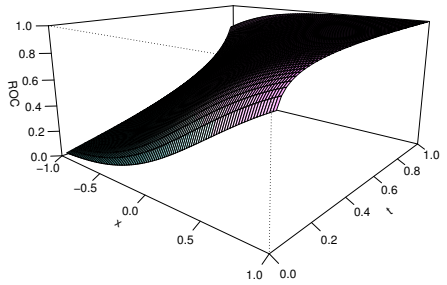
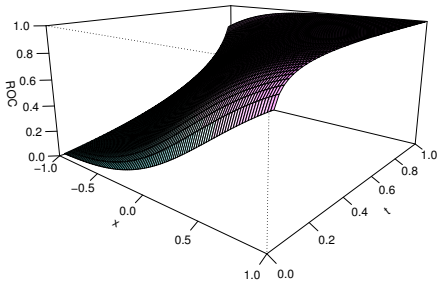


Figura 4.1: Boxplot funcional de las $\widehat{\text{AUC}}_x$ para $n = 100$ y $n = 200$ bajo el modelo (4.1) y (4.2). En verde se observa la verdadera AUC_x.

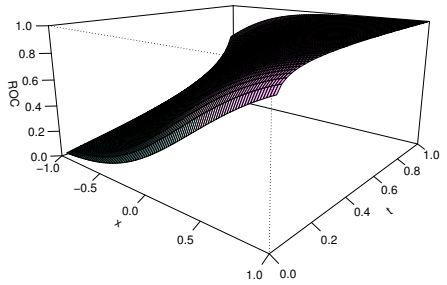
Además, en la Figura 4.2 se observa la superficie generada por las verdaderas curvas ROC bajo el modelo central y, en comparación, se presentan las superficies obtenidas con los estimadores clásicos y robustos a partir de una de las muestras generada para $n = 100$.



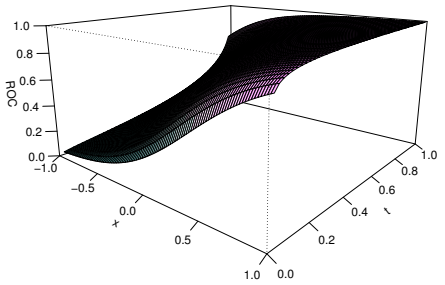
(a) Superficie ROC verdadera



(b) Estimadores Clásicos



(c) Estimadores Robustos 1



(d) Estimadores Robustos 2

Figura 4.2: Superficie ROC verdadera y superficies estimadas para $n_D = n_H = 100$ bajo el modelo (4.1) y (4.2).

Los gráficos y cuadros anteriores muestran que el desempeño de los estimadores propuestos es muy similar al del clásico en muestras sin contaminar.

Para analizar la sensibilidad de los estimadores de la curva ROC condicio-

nal ante desviaciones del modelo asumido, consideramos diferentes esquemas de contaminación introduciendo una porción de datos atípicos en ambas poblaciones. Propusimos contaminaciones del tipo corrimiento o desplazamiento (Shift) que consisten en sumar una cantidad fija a las muestras, variando la población afectada, el porcentaje de datos anómalos y el tamaño de la contaminación.

En todos los casos nos referiremos con C_0 a la situación en la que no se realiza ninguna contaminación y se consideran las muestras originales. A continuación detallamos los diferentes esquemas de contaminación:

- $C^H_{\delta,S}$: se contamina una proporción δ de la muestra de la población sana por *corrimiento*, sumándoles a las observaciones una cantidad $S\sigma_H$. De este modo, se reemplazan las primeras $m = \delta n$ observaciones de los individuos sanos siguiendo el modelo $Y_{H,i} = 0.5 + X_{H,i} + \sigma_H \epsilon_{H,i} + S\sigma_H$.
- $C^D_{\delta,S}$: se contamina una proporción δ de la población enferma por *corrimiento*, sumándoles a las observaciones una cantidad $S\sigma_D$. Aquí se reemplazan las primeras $m = \delta n$ observaciones de los individuos enfermos siguiendo el modelo $Y_{D,i} = 2 + 4X_{D,i} + \sigma_D \epsilon_{D,i} + S\sigma_D$.
- $C^{S,R}_{\delta}$: contaminación por *corrimiento* en la cual se suma una cantidad fija $S\sigma_H$ a los individuos sanos y $R\sigma_D$ a los enfermos, donde δ representa la proporción de cada población que se contamina. De este modo, las primeras $m = \delta n$ observaciones de cada población se reemplazan por observaciones generadas como:

$$Y_{H,i} = 0.5 + X_{H,i} + \sigma_H \epsilon_{H,i} + S\sigma_H,$$

$$Y_{D,i} = 2 + 4X_{D,i} + \sigma_D \epsilon_{D,i} + R\sigma_D.$$

Para las contaminaciones $C^H_{\delta,S}$ y $C^D_{\delta,S}$ consideramos $\delta = 0.05$ y 0.10, es decir, un $5\,\%$ y un $10\,\%$ de datos alterados. En relación al tamaño del corrimiento, tomamos S perteneciente al conjunto $\mathcal{S} = \{2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20\}$.

En las contaminaciones de tipo $C^{S,R}_{\delta}$ consideramos proporciones correspondientes a $\delta = 0.02$, 0.05 y 0.10 y para cada una de ellas analizamos las contaminaciones $C^{3.75,5}_{\delta}$, $C^{7.5,10}_{\delta}$ y $C^{15,20}_{\delta}$.

Para ejemplificar y mostrar el impacto de las contaminaciones directamente en las curvas ROC condicionales, en la Figura 4.3 se observan las superficies ROC estimadas bajo C_0 y bajo las contaminaciones $C^{15,20}_{0.10}$ a partir de una de las muestras generada con $n = 100$. En la imagen superior del panel derecho se puede ver que la superficie ROC estimada a partir la metodología clásica se distorsiona

notablemente ante la presencia de datos típicos. Por otro lado, para los métodos robustos se percibe una mayor estabilidad dado que las superficies estimadas son más parecidas entre sí.

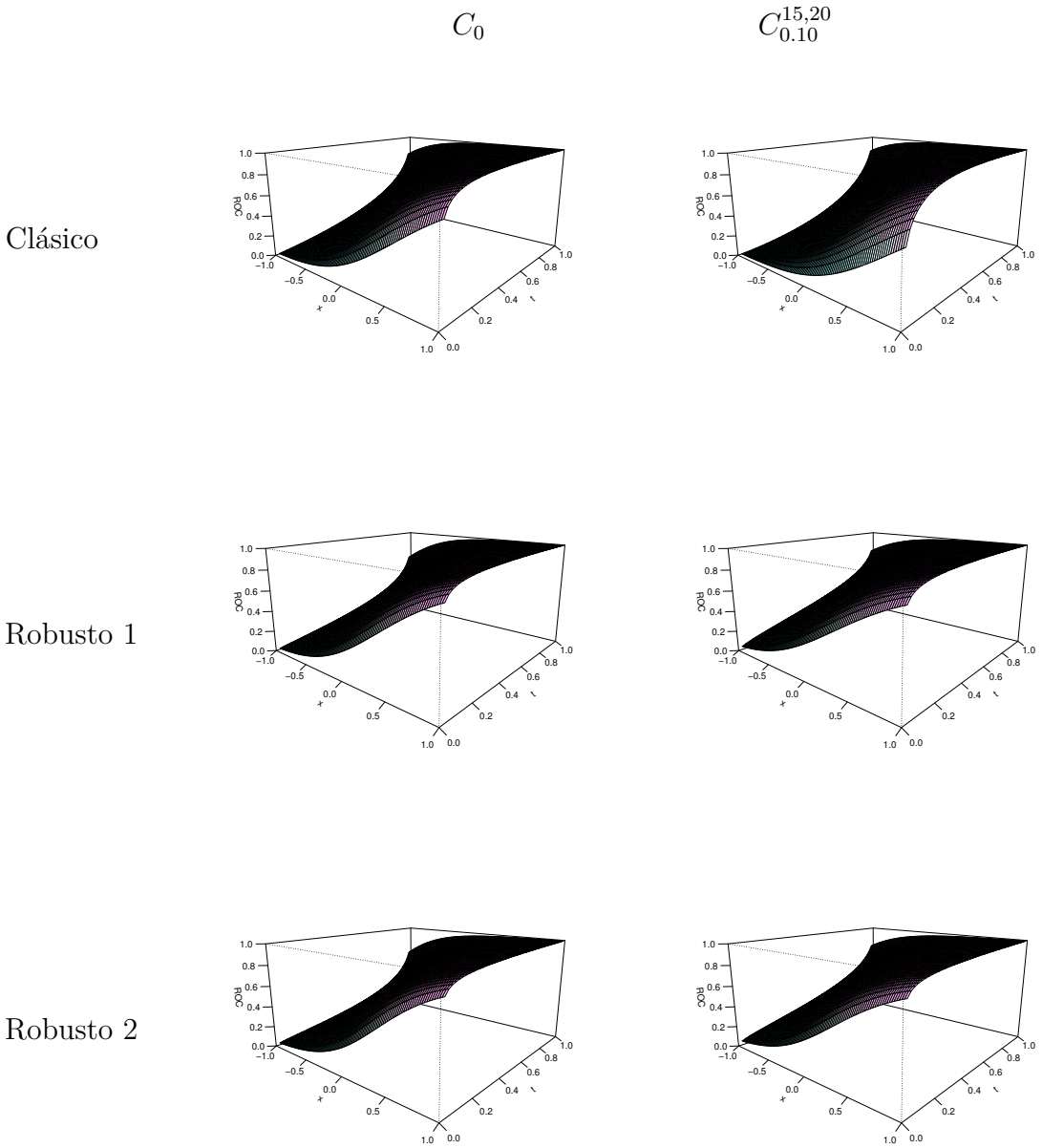


Figura 4.3: Superficies ROC estimadas para muestras limpias y contaminadas bajo $C_{0.10}^{15,20}$ con $n_D = n_H = 100$.

Los Cuadros 4.2 y 4.3 muestran para cada estimador los promedios obtenidos sobre las 1000 replicaciones de las medidas MSE y MAX bajo las contaminaciones del tipo $C_\delta^{S,R}$, para $n = 100$ y 200. Además, en la Figura 4.4 se presentan los boxplots correspondientes al MSE bajo C_0 y bajo las contaminaciones $C_\delta^{15,20}$ cuando el tamaño de las muestras es $n = 100$.

A partir de estos resultados, podemos ver que a medida que aumenta el porcentaje de contaminación y el tamaño de los corrimientos, las medidas MSE y MAX correspondientes al procedimiento clásico aumentan considerablemente. A modo de ejemplo, si analizamos el caso $n = 100$, vemos que bajo la contaminación $C_{0.02}^{15,20}$ el MSE aumenta un 48 % respecto a C_0. Al variar los porcentajes, tenemos que para $C_{0.05}^{15,20}$ la medida supera el doble del valor correspondiente a las muestras limpias, mientras que bajo $C_{0.10}^{15,20}$ es 4.6 veces mayor. En la medida MAX también vemos el impacto, si bien es un poco más moderado, dado que para $C_{0.02}^{15,20}$ y $C_{0.05}^{15,20}$ el aumento es del 24 % y 80 %, respectivamente, y para $C_{0.10}^{15,20}$ el MAX es 2.5 veces mayor que bajo C_0. En cambio, en los procedimientos robustos se observa que para las contaminaciones del 2 % y del 5 % el efecto es muy leve y las medidas se mantienen estables, aunque en el caso más extremo de $C_{0.10}^{15,20}$ se percibe cierto impacto, siendo que para $n = 100$ el MSE y el MAX aumentan un 60 % y un 20 %, respectivamente.

	Clásico		Robusto 1		Robusto 2	
Contaminación	MSE	MAX	MSE	MAX	MSE	MAX
C_0	0.0029	0.1300	0.0032	0.1372	0.0033	0.1412
$C_{0.02}^{3.75,5}$	0.0030	0.1325	0.0033	0.1391	0.0034	0.1429
$C_{0.02}^{7.5,10}$	0.0033	0.1390	0.0033	0.1383	0.0034	0.1421
$C_{0.02}^{15,20}$	0.0043	0.1611	0.0033	0.1387	0.0034	0.1422
$C_{0.05}^{3.75,5}$	0.0037	0.1632	0.0037	0.1462	0.0038	0.1496
$C_{0.05}^{7.5,10}$	0.0051	0.2008	0.0038	0.1451	0.0038	0.1483
$C_{0.05}^{15,20}$	0.0078	0.2350	0.0038	0.1467	0.0039	0.1497
$C_{0.10}^{3.75,5}$	0.0059	0.2351	0.0046	0.1637	0.0047	0.1663
$C_{0.10}^{7.5,10}$	0.0090	0.3009	0.0052	0.1647	0.0052	0.1669
$C_{0.10}^{15,20}$	0.0135	0.3347	0.0052	0.1646	0.0052	0.1668

Cuadro 4.2: Promedio de MSE y MAX sobre las replicaciones para cada contaminación por corrimiento del tipo $C_\delta^{S,R}$ cuando $n = 100$.

	Clásico		Robusto 1		Robusto 2	
Contaminación	MSE	MAX	MSE	MAX	MSE	MAX
C_0	0.0013	0.0868	0.0014	0.0906	0.0015	0.0940
$C_{0.02}^{3.75,5}$	0.0014	0.0926	0.0014	0.0920	0.0015	0.0948
$C_{0.02}^{7.5,10}$	0.0015	0.0986	0.0015	0.0920	0.0015	0.0946
$C_{0.02}^{15,20}$	0.0021	0.1145	0.0015	0.0922	0.0015	0.0947
$C_{0.05}^{3.75,5}$	0.0022	0.1370	0.0017	0.0978	0.0018	0.1000
$C_{0.05}^{7.5,10}$	0.0033	0.1756	0.0018	0.0990	0.0019	0.1007
$C_{0.05}^{15,20}$	0.0046	0.1958	0.0019	0.1000	0.0019	0.1015
$C_{0.10}^{3.75,5}$	0.0046	0.2273	0.0025	0.1186	0.0025	0.1200
$C_{0.10}^{7.5,10}$	0.0072	0.2925	0.0032	0.1254	0.0032	0.1259
$C_{0.10}^{15,20}$	0.0095	0.3129	0.0032	0.1254	0.0032	0.1258

Cuadro 4.3: Promedio de MSE y MAX sobre las replicaciones para cada contaminación por corrimiento del tipo $C_\delta^{S,R}$ cuando $n = 200$.

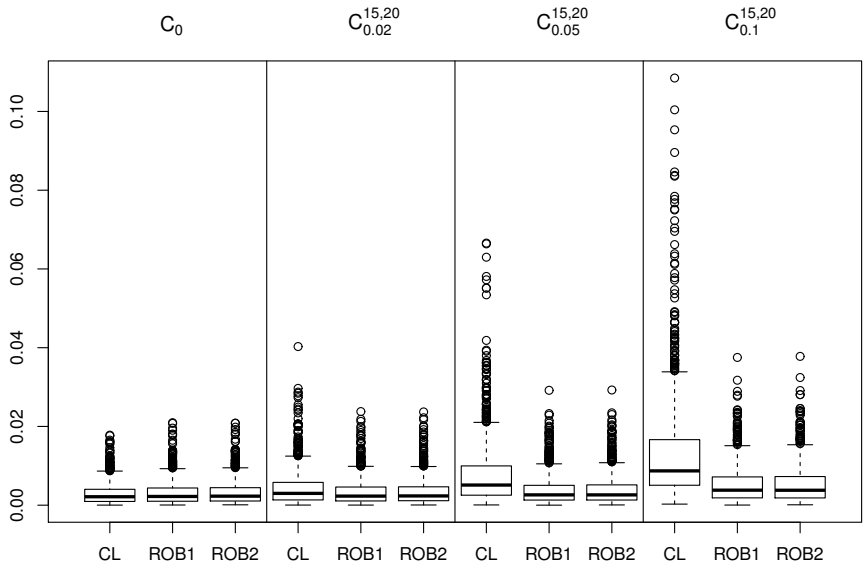


Figura 4.4: Boxplots de la medida MSE obtenidos sobre las 1000 replicaciones utilizando los estimadores clásicos y robustos bajo C_0 y $C_\delta^{15,20}$ cuando $n = 100$.

Este comportamiento también se puede apreciar en la Figura 4.4 donde se muestra que el boxplot del MSE para el método clásico no solo se desvía a medida que aumenta el porcentaje de contaminación, sino que además presenta una cantidad importante de valores atípicos. Por otra parte, en los estimadores robustos el impacto se percibe de forma mucho más moderada recién en la contaminación del 10 %.

Teniendo en cuenta el efecto observado en las medidas anteriores para los distintos estimadores, realizamos los boxplots funcionales de los estimadores $\widehat{\mathrm{AUC}}_x$ para evaluar y comparar el impacto de las mismas contaminaciones sobre el área bajo la curva condicional para los procedimientos clásicos y robustos. En las Figuras 4.5 y 4.6 podemos observar los resultados obtenidos cuando $n = 100$ y $n = 200$, bajo el esquema de contaminación $C_\delta^{S,R}$ con $S = 15$ y $R = 20$.

En los paneles izquierdos de las figuras se muestran los resultados obtenidos para el método de estimación clásica. En la contaminación $C_{0.02}^{15,20}$ se ve que, si bien la verdadera curva AUC_x se mantiene en el centro del boxplot funcional, ya es posible observar algunas curvas atípicas representadas en línea punteada en color rojo. A medida que aumentamos el porcentaje de contaminación observamos no solo un aumento de la cantidad de curvas atípicas, sino que las bandas del boxplot funcional son cada vez más anchas y, además, la verdadera AUC_x se desvía de la zona central de las curvas estimadas. En particular, en la contaminación $C_{0.10}^{15,20}$ y para valores de x mayores a cero, la curva verdadera se encuentra fuera de la región del 50 % central, tanto para $n = 100$ como para $n = 200$. Como era de esperar, corroborando el análisis realizado de los Cuadros 4.2 y 4.3, en los procedimientos robustos podemos ver estabilidad ante las contaminaciones $C_{0.02}^{15,20}$ y $C_{0.05}^{15,20}$, puesto que casi no se presentan curvas atípicas y la verdadera curva AUC_x se mantiene centrada en relación a las curvas estimadas. Para la contaminación $C_{0.10}^{15,20}$ se percibe un desvío de las curvas estimadas en relación a la curva verdadera, aunque este efecto es mucho más leve que en el método clásico.

En este punto resulta interesante destacar que los dos procedimientos de estimación robustos tienen un comportamiento muy similar, tanto bajo el modelo central como bajo el esquema de contaminación $C_\delta^{S,R}$. Este hecho refleja que el método de estimación Robusto 2 muestra cierta estabilidad frente a la especificación errónea del modelo en un sentido diferente al analizado a partir de la incorporación de datos atípicos, vinculado a los supuestos sobre la función de enlace. Siendo que el modelo considerado en (4.1) y (4.2) es binormal, la función link "natural" a tener en cuenta para la etapa correspondiente al modelo lineal generalizado es la función "probit". Por lo tanto, en el caso del estimador Robusto 2 observamos que asumiendo un modelo logístico (es decir, tomando una función link errónea) se están obteniendo resultados prácticamente iguales a los del estimador

Robusto 1, el cual supone apropiadamente un modelo normal.

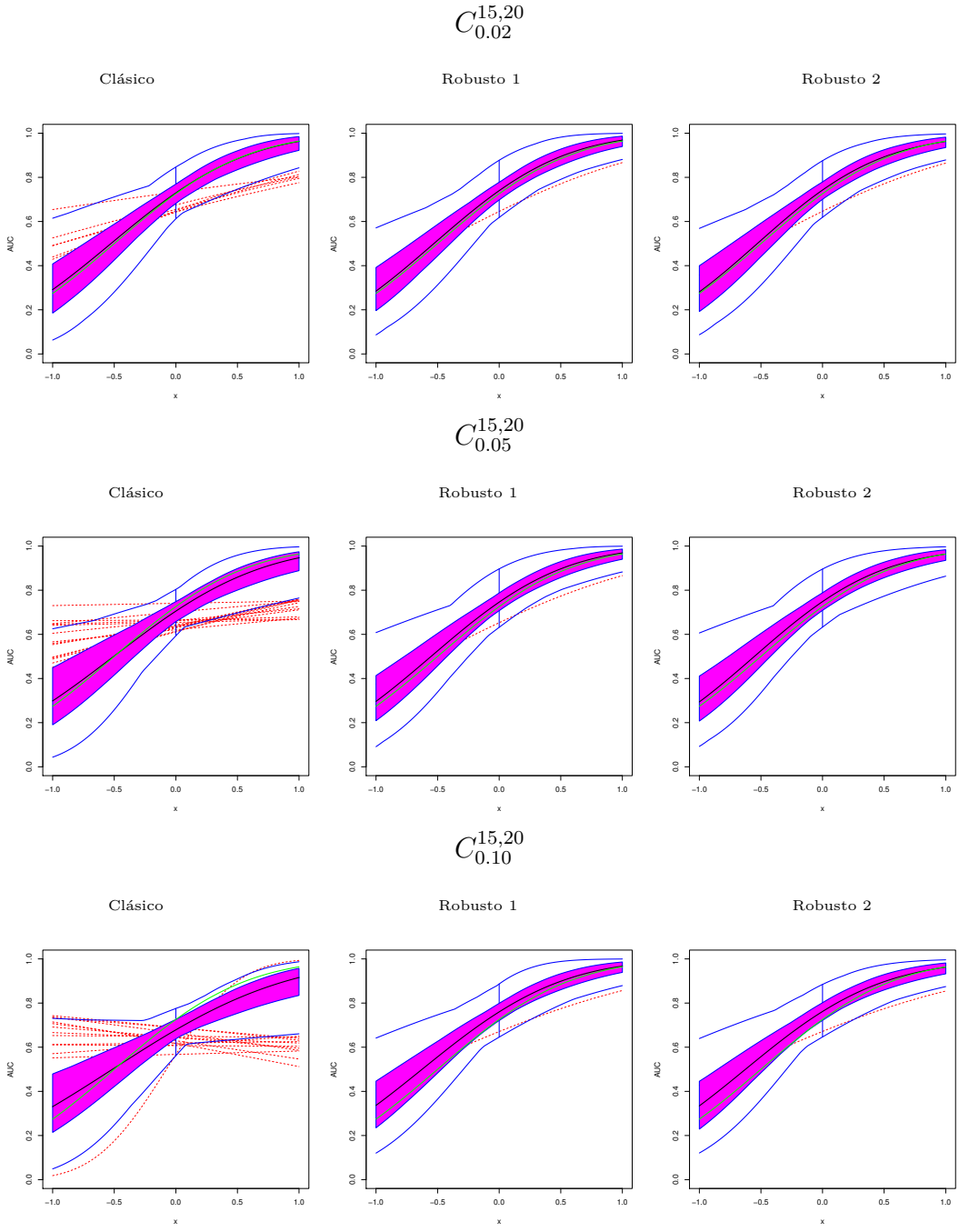


Figura 4.5: Boxplot funcional de las $\widehat{\text{AUC}}_x$ para el modelo (4.1) y (4.2), con $n = 100$, cuando se realizan las contaminaciones $C_{\delta}^{S,R}$ para $S = 15$ y $R = 20$. En verde se observa la verdadera AUC_x.

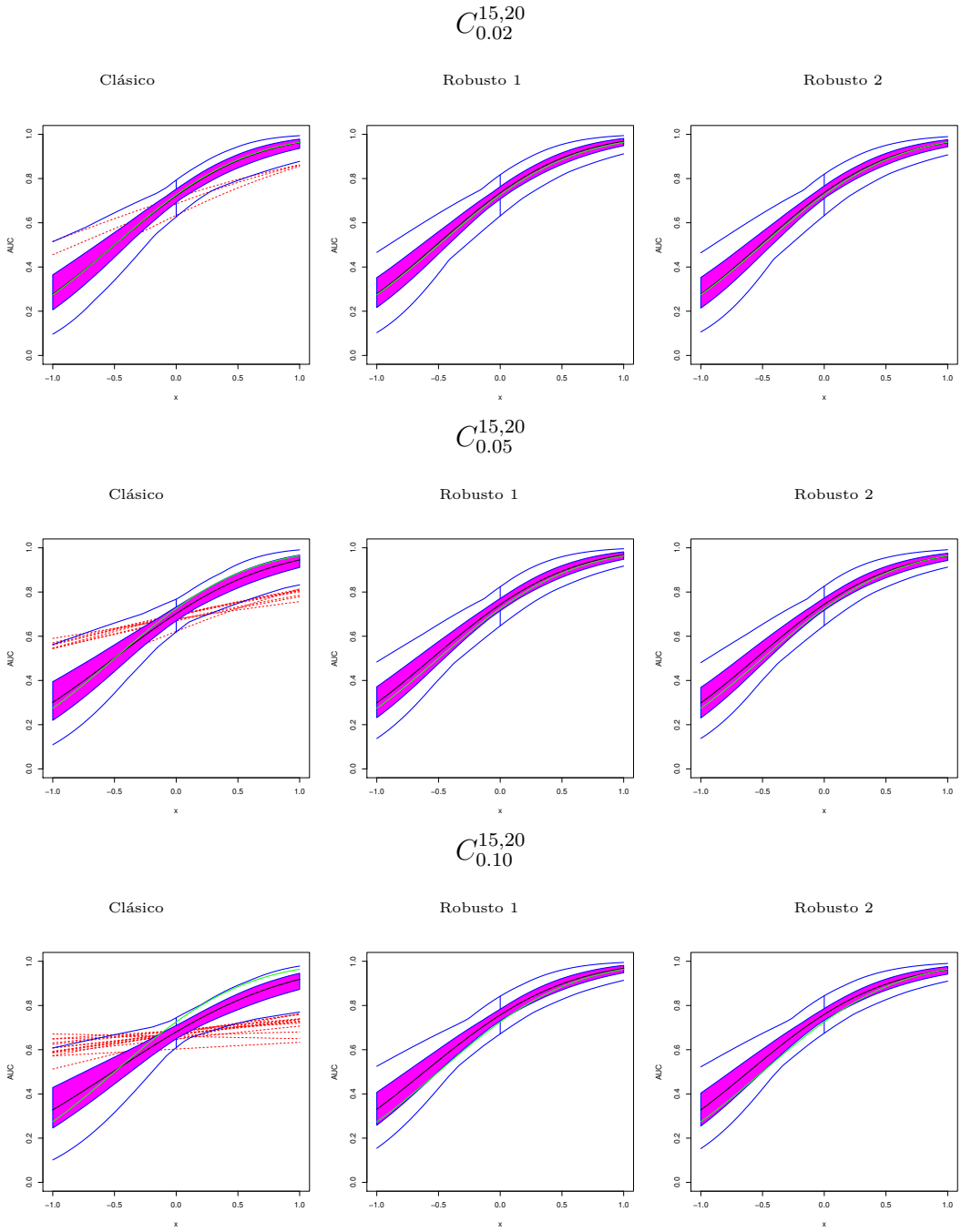


Figura 4.6: Boxplot funcional de las $\widehat{\text{AUC}}_x$ para el modelo (4.1) y (4.2), con $n = 200$, cuando se realizan las contaminaciones $C_{\delta}^{S,R}$ para $S = 15$ y $R = 20$. En verde se observa la verdadera AUC_x.

Para tener una mayor comprensión de la sensibilidad que tienen los métodos de estimación clásicos y robustos ante la presencia de datos atípicos, analizamos el efecto de diferentes contaminaciones en la población sana y en la población enferma por separado. En los Cuadros 4.4 y 4.5 se resumen los resultados de los promedios del MSE y MAX para las contaminaciones del 5 % y 10 % en cada una de las poblaciones cuando varía el tamaño del corrimiento, tomando muestras de tamaño $n_H = n_D = 100$. Como se puede apreciar, el procedimiento clásico se ve afectado significativamente a medida que se incrementa el tamaño del corrimiento S en la población sana, mientras que ante el mismo escenario de contaminación los métodos robustos se mantienen estables. Por ejemplo, bajo $C^H_{0.05,20}$ el MSE del estimador clásico es 4.5 veces mayor que el valor obtenido para muestras limpias y bajo $C^H_{0.10,20}$ es casi 10 veces más grande. Si comparamos con las contaminaciones mixtas realizadas previamente, vemos que este tipo de contaminación produce un efecto bastante más pronunciado.

En contraposición, se observa un impacto mucho más leve en el estimador clásico cuando se contamina un porcentaje de las observaciones de la población enferma. Notemos que bajo el escenario de contaminación $C^D_{0.10,S}$, con $S = 2.5$, el MSE ni siquiera llega a duplicarse en relación al valor correspondiente a C_0 y, a medida que se aumenta el tamaño de la contaminación, se mantiene prácticamente constante en ese mismo nivel. Para $C^D_{0.10,20}$ la medida MAX llega a aumentar un 25 % respecto del valor bajo C_0. A su vez, si miramos el comportamiento de los procedimientos robustos, vemos que presentan un desempeño prácticamente equivalente. Es decir, se puede notar cierto impacto de las contaminaciones cuando se trata de la población enferma, pero es un efecto moderado y similar al que se produce en el estimador clásico.

Este comportamiento se puede apreciar de forma mucho más clara en la Figura 4.7 donde se muestra el gráfico del MSE en función del nivel de corrimiento S, cuando $n = 100$, para los distintos porcentajes de contaminación. La línea roja corresponde al procedimiento de estimación clásica, la azul al Robusto 1 y la verde al Robusto 2. En el caso de las contaminaciones en la población sana, para el método clásico vemos cómo el MSE crece rápidamente a medida que varía el valor de S, mientras que para los métodos robustos la curva se mantiene estable. Por otra parte, para las contaminaciones en la población enferma la curva del MSE presenta un leve aumento al comienzo y luego se mantiene estable, de forma indistinta para todos los estimadores.

			S							
		C_0	2.5	5	7.5	10	12.5	15	17.5	20
			$C^H_{0.05,S}$							
	Cl	0.0029	0.0036	0.0057	0.0067	0.0075	0.0086	0.0099	0.0114	0.0131
MSE	Rob 1	0.0032	0.0035	0.0032	0.0032	0.0032	0.0032	0.0032	0.0032	0.0032
	Rob 2	0.0033	0.0036	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033
	Cl	0.1300	0.1557	0.2153	0.2360	0.2474	0.2601	0.2740	0.2886	0.3034
MAX	Rob 1	0.1372	0.1476	0.1379	0.1370	0.1379	0.1371	0.1370	0.1370	0.1370
	Rob 2	0.1412	0.1522	0.1419	0.1409	0.1417	0.1408	0.1407	0.1407	0.1407
			$C^D_{0.05,S}$							
	Cl	0.0029	0.0036	0.0037	0.0037	0.0037	0.0037	0.0037	0.0037	0.0037
MSE	Rob 1	0.0032	0.0038	0.0038	0.0038	0.0038	0.0038	0.0038	0.0038	0.0038
	Rob 2	0.0033	0.0038	0.0039	0.0039	0.0039	0.0039	0.0039	0.0039	0.0039
	Cl	0.1300	0.1398	0.1409	0.1410	0.1410	0.1410	0.1410	0.1410	0.1410
MAX	Rob 1	0.1372	0.1455	0.1463	0.1463	0.1463	0.1463	0.1463	0.1463	0.1463
	Rob 2	0.1412	0.1488	0.1495	0.1495	0.1495	0.1495	0.1495	0.1495	0.1495

Cuadro 4.4: Promedio de MSE y MAX para distintos valores de S en las contaminaciones $C^H_{0.05,S}$ y $C^D_{0.05,S}$, cuando $n = 100$.

			S							
		C_0	2.5	5	7.5	10	12.5	15	17.5	20
			$C^H_{0.10,S}$							
	Cl	0.0029	0.0065	0.0141	0.0163	0.0180	0.0202	0.0227	0.0255	0.0285
MSE	Rob 1	0.0032	0.0051	0.0033	0.0032	0.0032	0.0032	0.0032	0.0032	0.0032
	Rob 2	0.0033	0.0052	0.0034	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033
	Cl	0.1300	0.2244	0.3783	0.4040	0.4159	0.4280	0.4410	0.4543	0.4673
MAX	Rob 1	0.1372	0.1846	0.1419	0.1372	0.1371	0.1371	0.1371	0.1371	0.1371
	Rob 2	0.1412	0.1903	0.1461	0.1412	0.1412	0.1412	0.1412	0.1412	0.1412
			$C^D_{0.10,S}$							
	Cl	0.0029	0.0051	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053
MSE	Rob 1	0.0032	0.0051	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053
	Rob 2	0.0033	0.0051	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053	0.0053
	Cl	0.1300	0.1599	0.1635	0.1635	0.1635	0.1635	0.1635	0.1635	0.1635
MAX	Rob 1	0.1372	0.1634	0.1663	0.1663	0.1663	0.1663	0.1663	0.1663	0.1663
	Rob 2	0.1412	0.1659	0.1683	0.1684	0.1684	0.1684	0.1684	0.1684	0.1684

Cuadro 4.5: Promedio de MSE y MAX para distintos valores de S en las contaminaciones $C^H_{0.10,S}$ y $C^D_{0.10,S}$, cuando $n = 100$.

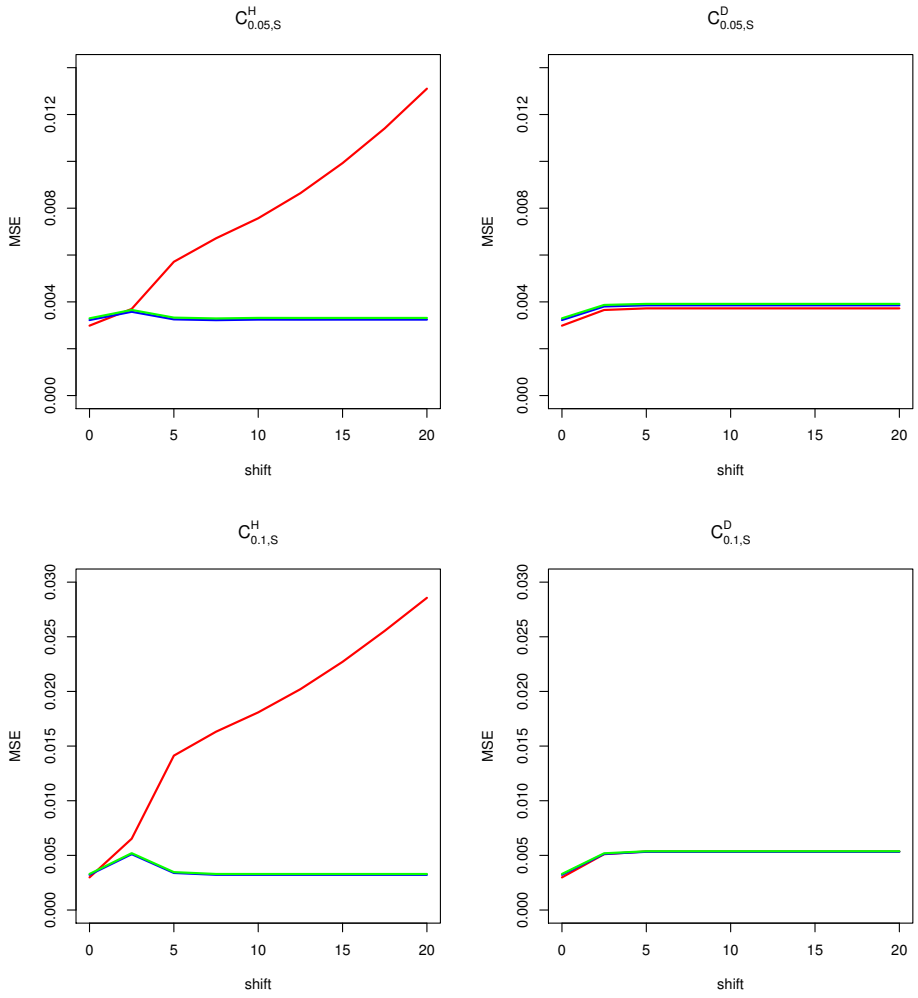


Figura 4.7: MSE en función del tamaño del corrimiento S bajo $C^H_{\delta,S}$ y $C^D_{\delta,S}$, para $\delta = 0.05, 0.10$ y $n_H = n_D = 100$. La línea roja corresponde al procedimiento de estimación clásica, la azul al Robusto 1 y la verde al Robusto 2.

Un efecto similar se ve reflejado en las Figuras 4.8 y 4.9 donde graficamos los boxplots funcionales de los estimadores del AUC_x bajo las contaminaciones $C^H_{\delta,S}$ y $C^D_{\delta,S}$, cuando $n_H = n_D = 100$, con pocentajes $\delta = 0.05, 0.10$ y distintos valores de S. En la Figura 4.8 vemos que para ambos porcentajes de contaminación en la población sana el corrimiento $S = 15$ causa un desvío notable en la estimación clásica del AUC_x. En cambio, como era de esperar, los estimadores robustos no presentan alteraciones importantes. Por otro lado, cuando contaminamos un 5 % de la población enferma con $S = 20$ no se observa un efecto de gran relevancia ni en el método clásico ni en los robustos, en tanto que para el 10 % la curva AUC_x

verdadera se desvía moderadamente respecto de las curvas estimadas (ver Figura 4.9), pero aún así se mantiene dentro del área magenta que representa el 50 % central de las curvas AUC_x.

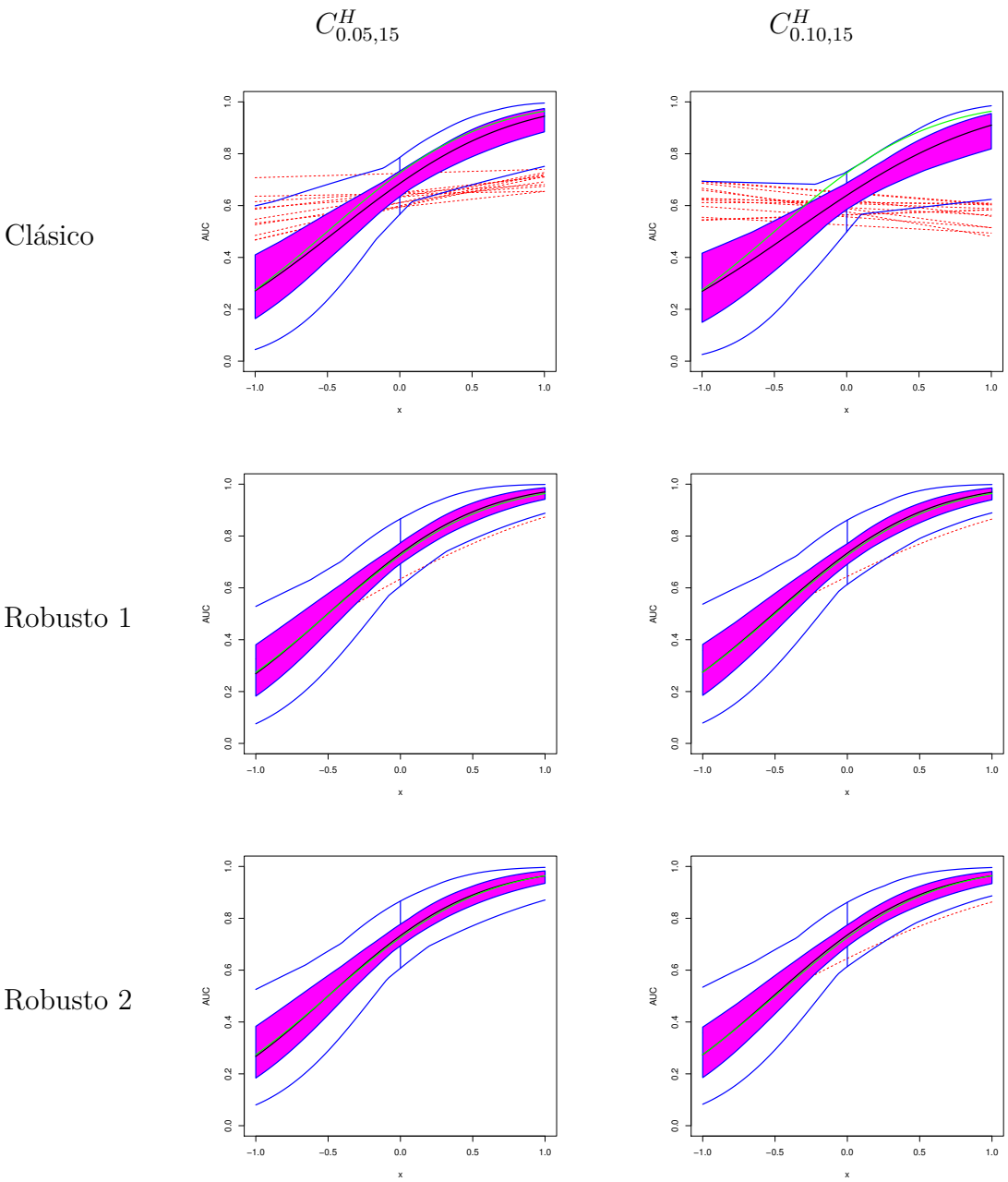


Figura 4.8: Boxplot funcional del $\widehat{\text{AUC}}_x$ para $n = 100$ bajo las contaminaciones $C^H_{\delta,S}$ cuando $S = 15$. En verde se observa la verdadera AUC_x.

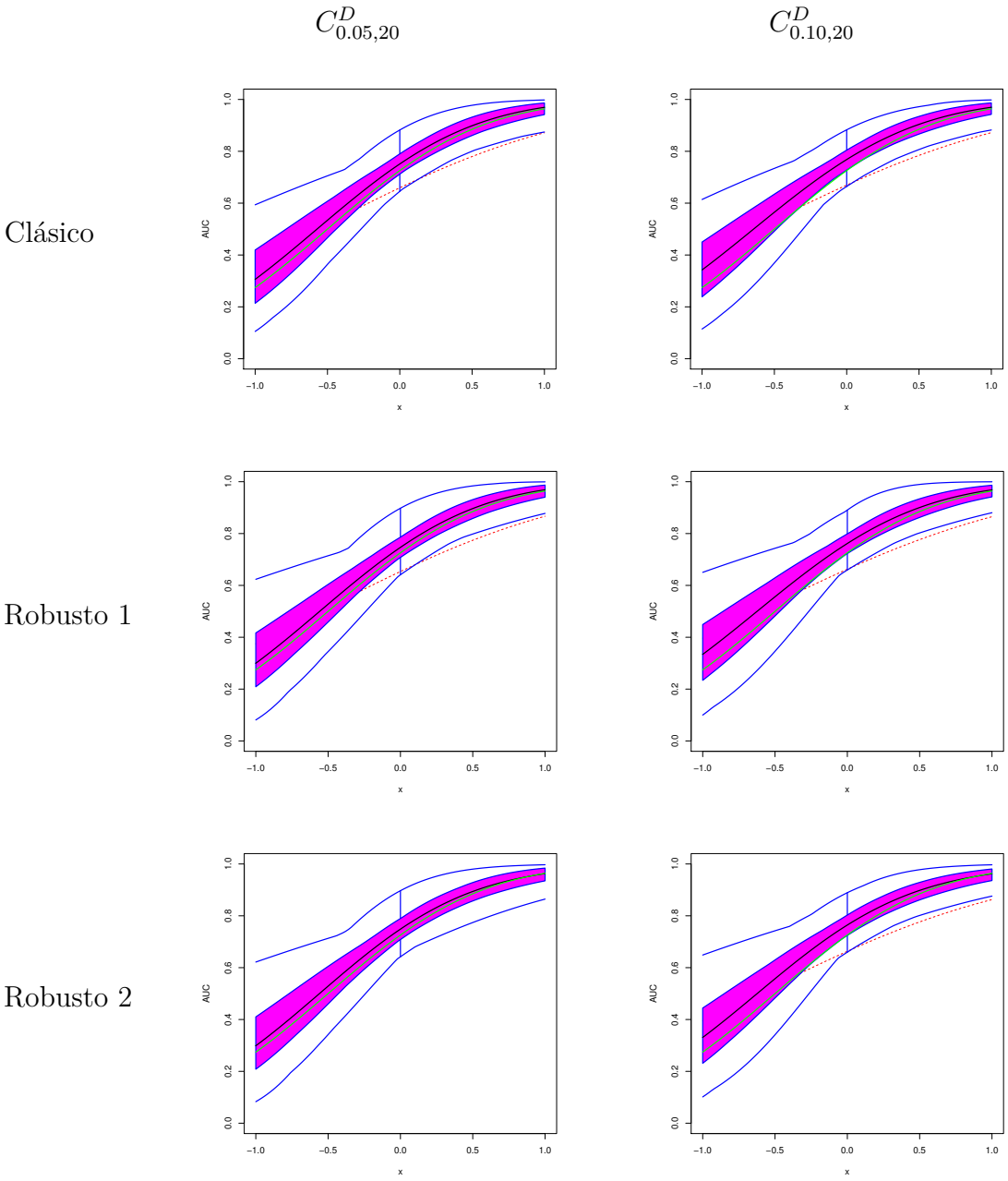


Figura 4.9: Boxplot funcional del $\widehat{\text{AUC}}_x$ para $n = 100$ bajo las contaminaciones $C^D_{\delta,S}$ cuando $S = 20$. En verde se observa la verdadera AUC_x.

Con el objetivo de profundizar en el análisis y tener una mayor comprensión de la diferencia percibida entre las contaminaciones en la población sana y la enferma, realizamos boxplots para los estimadores de los parámetros de las regresiones

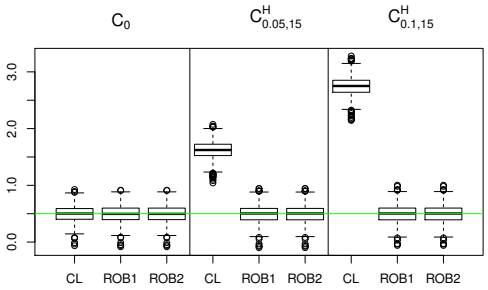
sobre las 1000 replicaciones, bajo los esquemas de contaminación $C^H_{\delta,15}$ y $C^D_{\delta,20}$ con $n_H = n_D = 100$ y $\delta = 0.05, 0.10$. En las Figuras 4.10 y 4.11 se muestran los boxplots correspondientes a $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\theta}_0$, $\widehat{\theta}_1$ y $\widehat{\alpha}_1$ para los distintos procedimientos de estimación y porcentajes de contaminación. Como se puede ver, los datos atípicos introducidos en la muestra sana provocan alteraciones evidentes en los estimadores clásicos de los parámetros. En algunos de ellos se observa un sesgo respecto del valor central, por ejemplo para β_0, mientras que en otros se puede apreciar un aumento de la variabilidad, como es el caso de las estimaciones del β_1. En cambio, cuando se contamina la población enferma los parámetros se ven prácticamente iguales en relación al caso C_0, de forma indistinta para todos los estimadores.

Vale la pena destacar que el comportamiento diferenciado del método Robusto 2 en los paneles c), d) y f) de las Figuras 4.10 y 4.11 que corresponden a los parámetros del modelo GLM se debe a que el modelo ajustado asume una función link logística en lugar de una probit.

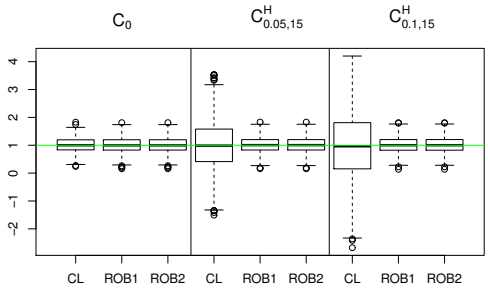
Luego del estudio realizado, pensamos que la diferencia entre el impacto de las contaminaciones en la población sana y en la enferma puede estar relacionada a las diversas etapas de estimación que tienen nuestros procedimientos y el rol distinguido que juegan las muestras en cada una de ellas. En las contaminaciones de la población sana, el efecto se produce directamente al comienzo en la estimación de los parámetros del modelo de regresión lineal y afecta a las etapas siguientes. Así, vemos cómo las propuestas robustas efectuadas en cada uno de los pasos logran estabilizar la situación. Sin embargo, el impacto de la muestra enferma aparece recién en una etapa bastante posterior, al momento de estimar los "placement values". Al parecer, el efecto de los datos atípicos introducidos no resulta tan directo en la construcción de la variable binaria, lo cual podría explicar la poca alteración que se visibiliza en la estimación de los parámetros del modelo lineal generalizado.

Para concluir, es interesante analizar qué ocurriría si se consideraran estimadores robustos solamente en las etapas de estimación correspondientes a los parámetros de las regresiones, es decir, si la estimación empírica de $\widehat{G}_H$ se realiza como en el método clásico. Para ilustrar dicho comportamiento, consideramos esta variante híbrida en la estimación del método Robusto 1. Calculamos mediante este procedimiento el promedio del MSE y MAX sobre las 1000 replicaciones y graficamos los boxplots funcionales del AUC cuando $n = 100$, bajo las contaminaciones en la población sana $C^H_{\delta,15}$ para porcentajes del 5 % y 10 %. Los resultados se presentan en el Cuadro 4.6 y en la Figura 4.12. Allí se puede observar que, si bien el impacto es más leve que en la estimación clásica, los datos atípicos introducidos producen un efecto considerable que se percibe en el aumento de las medidas MSE y MAX y en el desvío de las AUC, donde para el 10 % de contaminación vemos que la

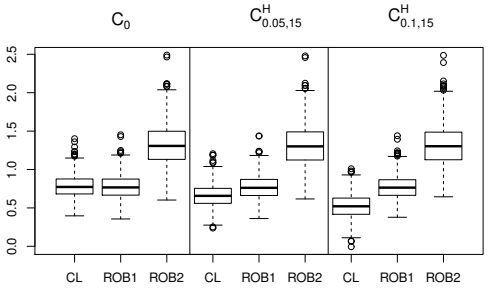
verdadera curva AUC queda por fuera de la banda del 50 % central de las curvas estimadas. Estos resulados dan prueba de que es necesario robustificar todos los pasos del proceso de estimación de las curvas ROC.



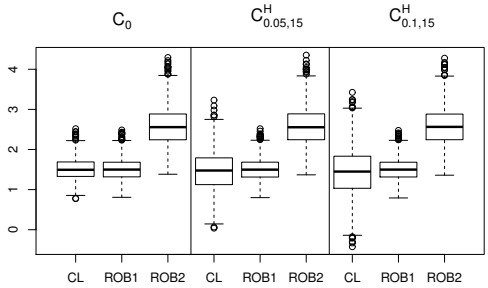
(a) Estimación de β_0. La línea verde corresponde al verdadero valor de β_0.



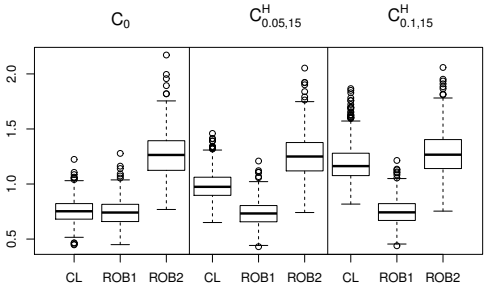
(b) Estimación de β_1. La línea verde corresponde al verdadero valor de β_1.



(c) Estimación de θ_0.

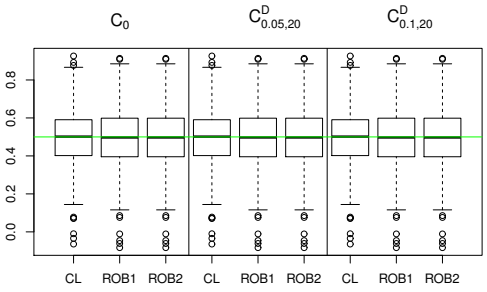


(d) Estimación de θ_1.

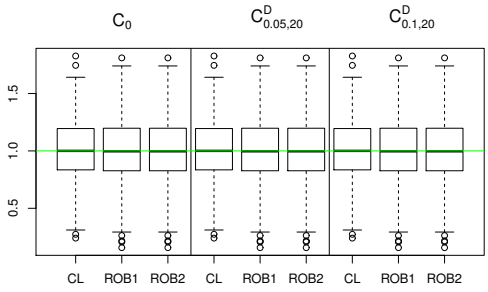


(e) Estimación de α_1.

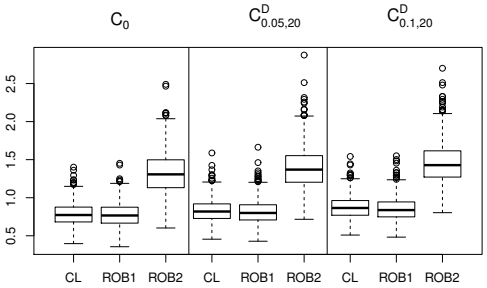
Figura 4.10: Boxplots de las estimaciones de los parámetros β_0, β_1, θ_0, θ_1 y α_1 sobre las 1000 replicaciones bajo C_0 y bajo las contaminaciones $C^H_{\delta,15}$ para $n_D = n_H = 100$.



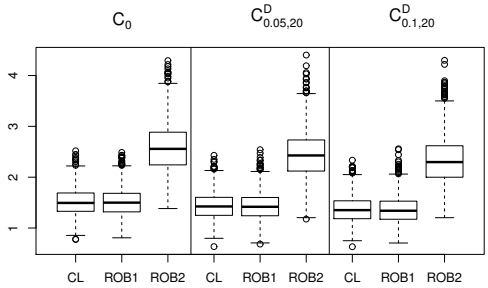
(a) Estimación de β_0. La línea verde corresponde al verdadero valor de β_0.



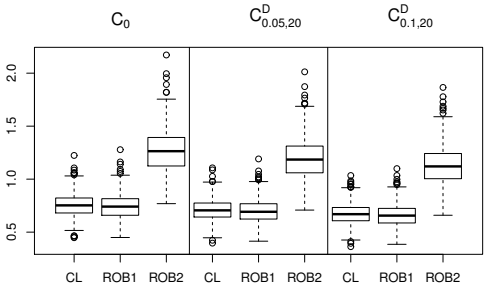
(b) Estimación de β_1. La línea verde corresponde al verdadero valor de β_1.



(c) Estimación de θ_0.



(d) Estimación de θ_1.

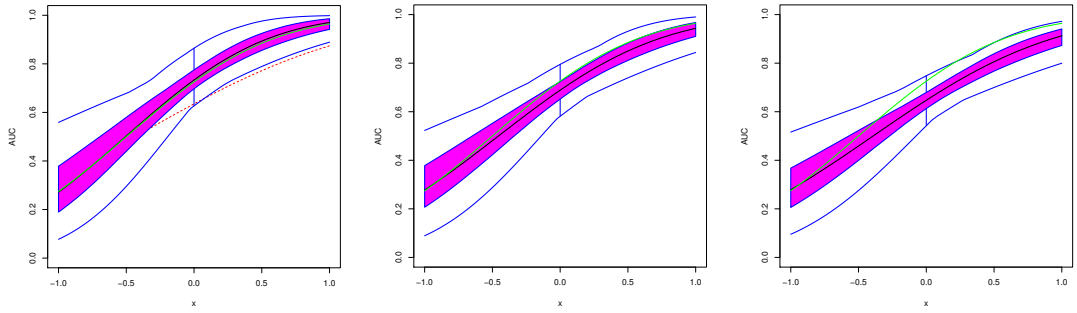


(e) Estimación de α_1.

Figura 4.11: Boxplots de las estimaciones de los parámetros β_0, β_1, θ_0, θ_1 y α_1 sobre las 1000 replicaciones bajo C_0 y bajo las contaminaciones $C^D_{\delta,20}$ para $n_D = n_H = 100$.

	C_0	$C^H_{0.05,15}$	$C^H_{0.10,15}$
MSE	0.0030	0.0059	0.0146
MAX	0.1324	0.2301	0.4050

Cuadro 4.6: Promedio de MSE y MAX obtenido para el procedimiento híbrido cuando $n = 100$, bajo las contaminaciones $C^H_{\delta,S}$ para $S = 15$



C_0 $C^H_{0.05,15}$ $C^H_{0.10,15}$

Figura 4.12: Boxplot funcional del $\widehat{\text{AUC}}_x$ obtenido para el procedimiento híbrido con $n = 100$, bajo las contaminaciones $C^H_{\delta,S}$ cuando $S = 15$. En verde se observa la verdadera AUC_x.

Capítulo 5

Aplicación a datos reales: Identificación de discapacidad auditiva durante el período neonatal

En este capítulo realizaremos un análisis desde una perspectiva robusta a partir de un conjunto de datos estudiados en Norton *et al.* (2000), sobre distintos biomarcadores para identificar problemas auditivos en recién nacidos.

Exploraremos, por una parte, el efecto en la capacidad discriminatoria de los biomarcadores cuando se contemplan covariables como la edad y el sexo de los sujetos. Por otro lado, el objetivo central consiste en analizar la potencialidad de las metodologías robustas propuestas en nuestro trabajo dentro de este contexto de aplicación, ante el posible impacto de datos atípicos en la estimación de las curvas ROC condicionales.

5.1. Conjunto de datos

El conjunto de datos en el cual nos basamos proviene de un estudio de audiología neonatal desarrollado en Norton *et al.* (2000). Dicho estudio fue diseñado con el propósito de evaluar la capacidad de tres biomarcadores (DPOAE, TEOAE y ABR) para identificar problemas auditivos en niños recién nacidos. Se consideró una selección de bebés a los que se les realizaron los tres tests durante el período neonatal y, posteriormente, entre los 8 y los 12 meses de edad se determinó quiénes

presentaban discapacidad auditiva tomando como criterio las mediciones de VRA ("Visual Reinforcement Audiometry"). Junto con el estado de audición, se registró también la edad que tenían en dicho momento y el sexo.

Los datos disponibles [1] contienen 5056 observaciones de mediciones de los tres biomarcadores tomadas en el oído izquierdo y en el derecho. Dentro de este conjunto de datos, para nuestro estudio consideramos los resultados del biomarcador ABR medido en el oído izquierdo. Así, contamos con un total de 2540 observaciones entre las cuales 80 corresponden a individuos clasificados con discapacidad auditiva (considerados enfermos) y 2460 a individuos con audición normal (considerados sanos).

5.2. Análisis de los datos

De ahora en adelante, notaremos con $Y_{D,i}$ y $\boldsymbol{X}_{D,i}$, $1 \leq i \leq 80$, al biomarcador ABR y a las covariables medidas en los individuos enfermos y con $Y_{H,i}$ y $\boldsymbol{X}_{H,i}$, $1 \leq i \leq 2460$, a las correspondientes en los individuos sanos. Como mencionamos previamente, las covariables consideradas fueron la edad (E) de los bebés al momento de determinar el estado de audición y el sexo (S), donde la primera se mide en semanas de vida y la segunda está representada por una variable binaria (de valor 0 para el sexo femenino y 1 para el masculino). De este modo, identificaremos a las covariables como vectores independientes $\boldsymbol{X}_{D,i} = (E_{D,i}, S_{D,i})$ y $\boldsymbol{X}_{H,i} = (E_{H,i}, S_{H,i})$.

Estimamos las curvas ROC condicionales y sus respectivas AUC aplicando las metodologías de estimación clásica y robusta descriptas en los capítulos anteriores. Siguiendo el análisis llevado a cabo por Janes, Longton y Pepe (2009), asumimos un modelo lineal para el biomarcador en la población sana tal que

$$Y_{H,i} = \beta_0 + \beta_1 E_{H,i} + \beta_2 S_{H,i} + \sigma_H \epsilon_{H,i}, \tag{5.1}$$

para todo $1 \leq i \leq 2460$.

La curva ROC condicional fue modelada como

$$\mathrm{ROC}_{\boldsymbol{X}}(t) = g\left(\theta_0 + \theta_1 E + \theta_2 S + h(t)\right),$$

con $h(t) = \alpha_1 \phi^{-1}(t)$, siendo ϕ la función de distribución acumulada de una variable aleatoria normal estándar. En ambas metodologías elegimos como función link g la función "probit", es decir, $g = \phi$.

[1] Los datos se pueden descargar desde el sitio: `https://research.fredhutch.org/diagnostic-biomarkers-center/en/datasets.html`

5.2.1. Metodología de estimación directa clásica

A continuación se presentan los resultados de la estimación clásica de las curvas ROC condicionales, denotadas por $\widehat{\text{ROC}}_{\boldsymbol{X}}(t)$, y de las $\widehat{\text{AUC}}_{\boldsymbol{X}}$ asociadas.

Para evaluar los estimadores se utilizó una grilla en t de 49 puntos equidistantes tomados entre 0.02 y 0.98. En relación a la variable edad E, contemplamos una grilla de 121 puntos con valores entre 29 y 53 (semanas) y se calcularon los estimadores para cada una de esas edades y cada valor de sexo $S = 0$ y $S = 1$.

En la Figura 5.1 se observan las curvas ROC obtenidas con los estimadores clásicos condicionadas al sexo femenino y masculino para distintas edades, tomando $E = 31, 38, 45$ y 51. A su vez, la Figura 5.2 muestra las superficies ROC estimadas sobre las grillas de t y de edades para los valores de la variable sexo 0 y 1. Como se puede apreciar, las curvas ROC varían considerablemente a medida que aumenta la edad, mejorando la capacidad discriminatoria del biomarcador. Por otra parte, a simple vista, la diferencia entre las curvas correspondientes al sexo femenino y al masculino no parece ser tan pronunciada.

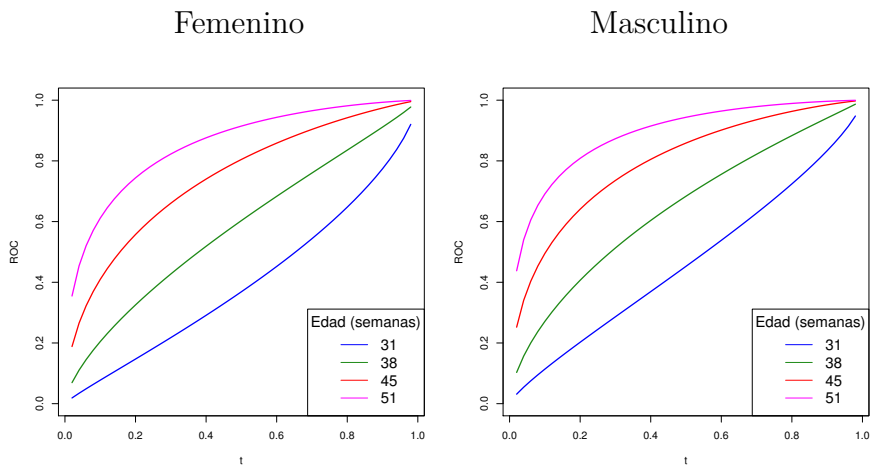


Figura 5.1: Curvas $\widehat{\text{ROC}}_{\boldsymbol{X}}(t)$ condicionadas al sexo femenino y masculino para distintos valores de la covariable edad: 31 (azul), 38 (verde), 45 (rojo) y 51 (magenta).

Femenino Masculino

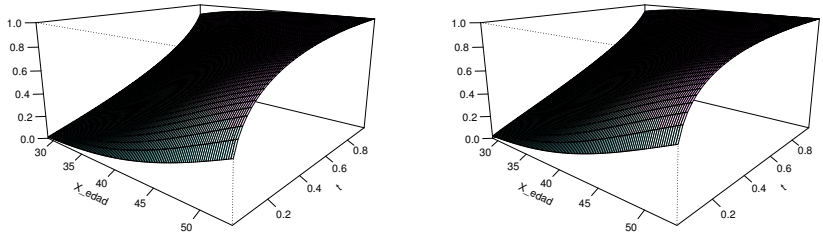


Figura 5.2: Superficies ROC estimadas con la metodología directa clásica para el sexo femenino y masculino, en base a diferentes edades y valores de t.

Para tener una mayor comprensión, en la Figura 5.3 representamos en un mismo gráfico las curvas AUC estimadas para ambos sexos, en función de las distintas edades. Allí se puede ver que en los dos casos el valor del AUC crece a medida que la edad aumenta. Además, queda en evidencia que la capacidad discriminatoria del test resulta mejor para el sexo masculino que para el femenino, ya que el área bajo la curva asociada es mayor en todas la edades consideradas.

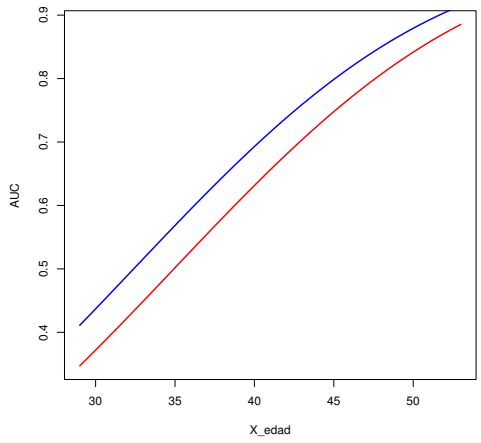


Figura 5.3: Curvas AUC estimadas según el sexo, para los distintos valores de edad entre 29 y 53 semanas. En rojo se representa la curva correspondiente al sexo femenino y en azul la del sexo masculino.

5.2.2. Datos atípicos y estimación robusta

Antes de evaluar los resultados de los estimadores clásicos presentados anteriormente y realizar una comparación con los estimadores robustos, exploramos, primero, la existencia de observaciones anómalas en la muestra que pudieran estar afectando a las estimaciones de las curvas ROC condicionales.

Para eso, nos basamos en los residuos del ajuste robusto del modelo de regresión lineal propuesto en (5.1) para la población sana, computado a partir de un MM-estimador como el que describimos en los capítulos anteriores. De este modo, detectamos la presencia de 309 datos atípicos que representan aproximadamente el 12 % de los individuos de la muestra sana. En la Figura 5.4 se muestra el boxplot de los residuos obtenidos para dicho modelo lineal, donde se pueden ver las observaciones cuyos residuos se encuentran alejados de la mayoría de los datos.

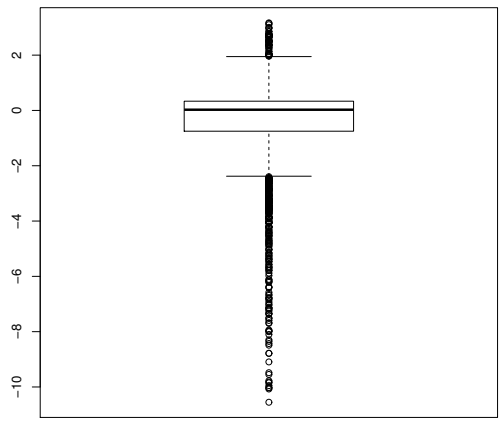


Figura 5.4: Boxplot de los residuos del ajuste robusto en el modelo de regresión lineal para el biomarcador ABR en los individuos sanos.

Teniendo en mente el comportamiento que observamos en nuestro estudio de simulación ante las contaminaciones en la población sana y que en los datos identificamos un porcentaje significativo de observaciones atípicas, computamos, entonces, los estimadores robustos de las curvas ROC condicionales. Presentaremos únicamente los resultados obtenidos a partir del procedimiento de estimación denominado “Robusto 1” en el Capítulo 4, puesto que las dos metodologías robustas estudiadas tienen un desempeño similar.

En la Figura 5.5 representamos, para cada sexo por separado, las curvas $\widehat{\text{ROC}}_{\boldsymbol{X}}(t)$ estimadas a partir de la metodología robusta, evaluadas en las diferentes edades seleccionadas ($E = 31, 38, 45, 51$). A modo comparativo, en línea discontinua se muestran las estimaciones clásicas de las curvas ROC condicionales correspondientes. De este modo, podemos apreciar diferencias notorias entre ambos estimadores, particularmente para edades más bajas donde vemos que la capacidad discriminatoria del test parece ser menor de acuerdo a las estimaciones robustas.

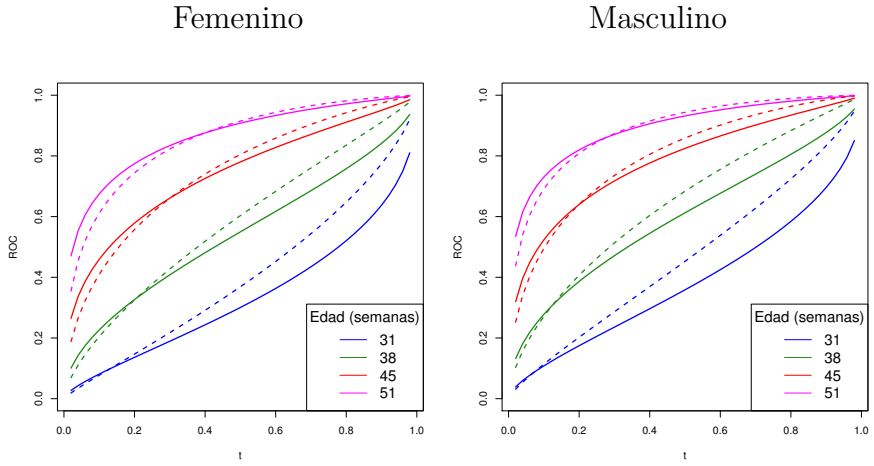


Figura 5.5: Curvas $\widehat{\text{ROC}}_{\boldsymbol{X}}(t)$ condicionadas al sexo femenino y masculino para distintos valores de la covariable edad. La estimación robusta está representada en línea continua y en línea discontinua se grafican las estimaciones clásicas correspondientes.

Esta última característica se ve también reflejada en la Figura 5.6 donde graficamos las curvas $\widehat{\text{AUC}}_{\boldsymbol{X}}$, según el sexo, para ambas metodologías de estimación. Allí observamos, por ejemplo, que para el sexo masculino la curva AUC estimada de forma robusta se encuentra por debajo de la clásica en todas las edades menores a 50 semanas. Algo similar sucede para el sexo femenino en las edades menores a aproximadamente 47 semanas. En cambio, para los individuos que se encuentran en las edades mayores el estimador robusto indicaría un desempeño levemente mejor que el clásico.

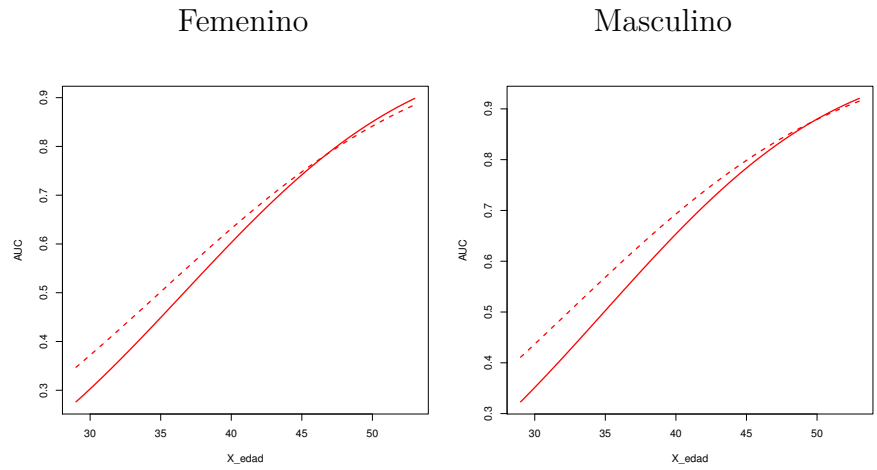


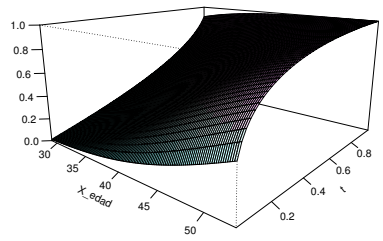
Figura 5.6: Curvas AUC estimadas según el sexo, para los distintos valores de edad entre 29 y 53 semanas. En línea continua se grafican los resultados del estimador robusto y en línea discontinua los del clásico.

Para tener una referencia en el análisis, además de estimar las curvas ROC condicionales mediante la propuesta robusta, calculamos nuevamente el estimador clásico quitando las 309 observaciones atípicas.

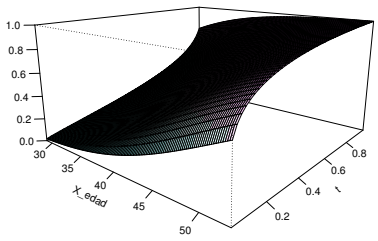
La Figura 5.7 muestra las superficies ROC estimadas para el sexo femenino con las metodologías clásica y robusta, tomando la muestra completa, y con el estimador clásico pero para la nueva muestra sin los datos atípicos. En todos los casos la grilla utilizada es la misma que la descripta en la Sección 5.2.1.

A su vez, en la Figura 5.8 se presentan en paralelo los resultados de las curvas $\widehat{\mathrm{ROC}}_{\boldsymbol{X}}(t)$ obtenidas con esos mismos tres mecanismos de estimación para ambos sexos y distintas edades. En dicha figura se pone en evidencia que el estimador clásico computado sin los datos atípicos y el robusto con los datos completos son bastante similares entre sí, mientras que el estimador clásico calculado con la muestra original se distingue en una buena parte de los valores de t, cualquiera sea la edad que se mire.

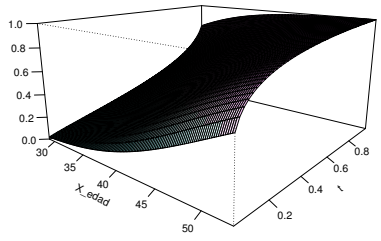
Asimismo, este comportamiento se ve plasmado en la Figura 5.9 donde representamos las gráficas de las curvas $\widehat{\mathrm{AUC}}_{\boldsymbol{X}}$ condicionadas al sexo femenino y masculino para los distintos estimadores.



(a) Estimación Clásica



(b) Estimación Robusta



(c) Estimación Clásica sin datos atípicos

Figura 5.7: Superficies ROC condicionadas al sexo femenino para el estimador clásico y robusto con todos los datos, y para el clásico sin las observaciones atípicas.

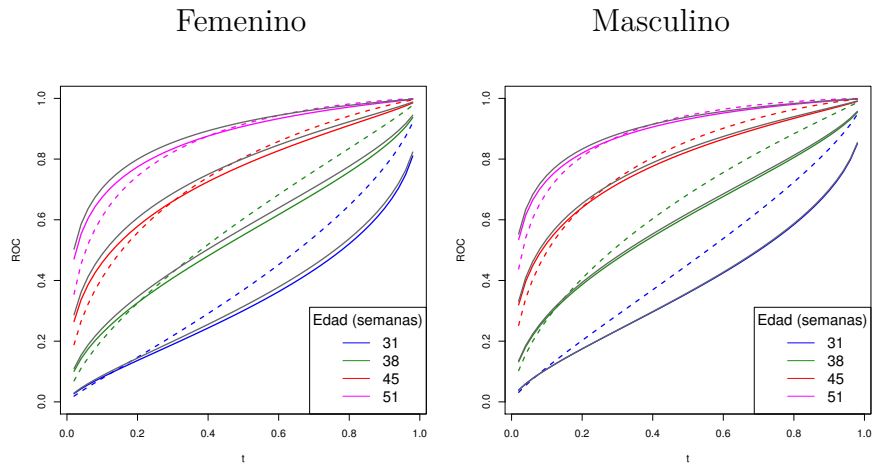


Figura 5.8: Curvas $\widehat{\text{ROC}}_{\boldsymbol{X}}(t)$ condicionadas al sexo femenino y masculino para distintas edades. En colores, las estimaciones robustas (línea continua) y clásicas (línea discontinua) con todos los datos involucrados y, en gris, las estimaciones clásicas para la muestra sin datos atípicos (línea en color gris).

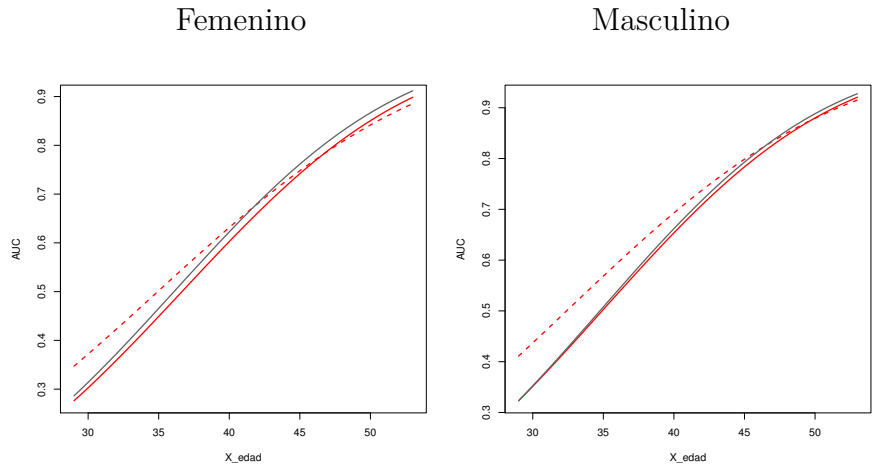
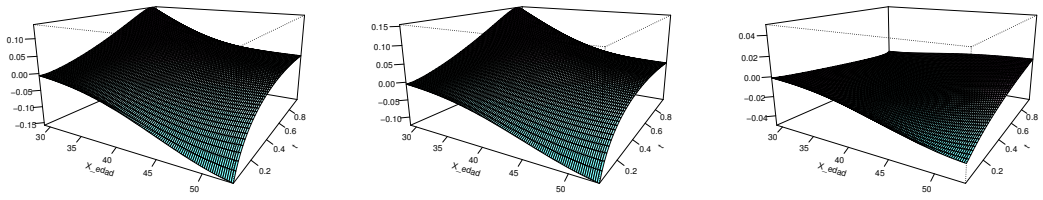


Figura 5.9: Curvas AUC estimadas según el sexo, en función de la edad. En rojo se grafican los resultados del estimador robusto (línea continua) y del clásico (línea discontinua) cuando se consideran todos los datos. En línea gris observamos los resultados del estimador clásico cuando se eliminan las observaciones atípicas.

Por último, para tener una perspectiva más clara en la comparación de estos tres estimadores, en la Figura 5.10 se muestran las diferencias entre las superficies ROC estimadas con cada uno de ellos. Como podemos apreciar, cuando se compara el método clásico, calculado a partir de la muestra completa, con el mismo sin los

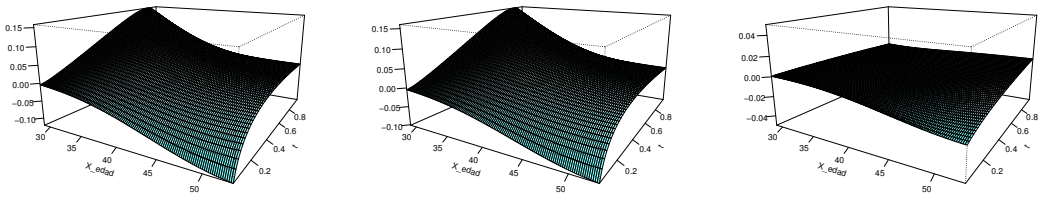
datos atípicos y con el método robusto, las diferencias alcanzan valores en módulo de aproximadamente 0.15. En cambio, si miramos la diferencia entre la superficie estimada con la propuesta clásica sin los datos atípicos y la correspondiente al estimador robusto en base a la muestra completa, los valores llegan a niveles mucho menores, variando aproximadamente entre -0.03 y 0.01.

Femenino



(a) (b) (c)

Masculino



(a) (b) (c)

Figura 5.10: Diferencias entre las superficies ROC estimadas: (a) Estimador clásico con los datos completos vs. sin datos atípicos. (b) Estimador clásico vs. robusto, con todos los datos involucrados. (c) Estimador robusto con los datos completos vs. estimador clásico sin los datos atípicos.

De esta forma, teniendo en cuenta todo el análisis realizado, podemos concluir que las observaciones atípicas influyen en los estimadores clásicos, aumentando o disminuyendo la capacidad discriminatoria del biomarcador ABR en función de las edades que se consideren, y que la metodología robusta muestra un desempeño más estable que la clásica.

Conclusiones

En este trabajo nos enfocamos en el modelado de la curva ROC, la cual representa una herramienta potente para evaluar la capacidad de un biomarcador de distinguir entre dos clases o estados. En particular, estudiamos la estimación de curvas ROC en presencia de covariables, teniendo en cuenta que resulta conveniente incluir la información adicional que proveen en el análisis ya que pueden afectar a la efectividad y a la capacidad discriminatoria del biomarcador.

Dentro del contexto de modelos de regresión para curvas ROC condicionales, revisamos la metodología directa en la cual se propone un modelo para la curva ROC en sí misma y los efectos de las covariables se evalúan directamente. Abordamos un método semiparamétrico que combina estimadores clásicos para modelos de regresión paramétricos con un estimador basado en la función de distribución empírica.

Teniendo en consideración que los estimadores involucrados en las diversas etapas de la metodología de regresión directa pueden verse muy afectados cuando los datos contienen observaciones anómalas, realizamos un experimento numérico para evaluar la sensibilidad del método ante un porcentaje de datos atípicos en la muestra. En dicho experimento analizamos contaminaciones por separado en las muestras correspondendientes a las poblaciones enferma y sana y, a su vez, contaminaciones en ambas poblaciones en simultáneo. El estudio revela que los datos atípicos provocan un efecto importante en el estimador de la curva ROC condicional y en sus medidas de resumen. Sin embargo, también se observa que las contaminaciones estudiadas en la población enferma tienen un impacto mucho más leve en comparación con los otros esquemas de contaminación. Esto parece estar relacionado a que las muestras de enfermos y sanos no juegan un rol equivalente en los distintos pasos de estimación requeridos en la metodología directa.

A su vez, en nuestro trabajo propusimos un procedimiento robusto para estimar la curva ROC condicional a partir de una adaptación del algoritmo presentado en la metodología directa clásica. Para eso, previamente realizamos una revisión de distintos enfoques de estimación robusta tanto en el marco paramétrico como en el

no paramétrico. En el método propuesto se incorporan estimadores robustos para modelos de regresión combinados con una función de distribución empírica pesada que amortigua el efecto de los residuos grandes en la estimación de la distribución de la población sana.

A partir del estudio de simulación analizamos el desempeño del procedimiento robusto sobre las muestras originales y las contaminadas, al igual que con el método directo clásico. Los resultados muestran que el estimador propuesto se mantiene estable y, en general, resulta mucho más resistente que el estimador clásico, salvo en las contaminaciones de la población enferma que presentan un comportamiento similar.

En este sentido, para futuras investigaciones sería interesante explorar otro tipo de contaminaciones en la muestra de la población enferma que provoquen una desestabilización mayor en el estimador clásico y, de este modo, profundizar en los alcances del método robusto. A su vez, a partir de nuestro trabajo se podría continuar estudiando otros escenarios para las distribuciones del biomarcador en las poblaciones enferma y sana, y extender la propuesta a otro tipo de modelos.

Por último, en relación a la aplicación de estos estimadores, realizamos un estudio preliminar en un ejemplo de audiología neonatal, en el cual analizamos un biomarcador que se utiliza para la identificación de problemas auditivos. Por una parte, los resultados muestran diferencias notorias en el desempeño del biomarcador cuando se tienen en cuenta las covariables disponibles, lo que permite indagar en las condiciones bajo las cuales se debe aplicar el test y mejorar la interpretación de sus resultados. Por otro lado, identificamos una porción de datos atípicos en la muestra y estudiamos el impacto que provocan en la estimación de las curvas ROC. Como era de esperar, vimos que dichas observaciones tienen influencia en la propuesta de estimación clásica y concluimos que el procedimiento robusto presenta un desempeño más confiable, tomando como referencia los resultados obtenidos a partir del procedimiento de estimación clásica cuando se eliminan los datos anómalos de la muestra.

Bibliografía

[1] Analytical Methods Committee. Robust statistics – How not to reject outliers. *Analyst 114* (1989), 1693–1702.

[2] Bianco, A. M., Boente, G., y González-Manteiga, W. A robust approach for ROC curves with covariates. *arXiv:2007.00150 [stat.ME]* (2020). Disponible en: `https://doi.org/10.48550/arXiv.2007.00150`.

[3] Bianco, A. M., y Yohai, V. J. Robust estimation in the logistic regression model. *In: Rieder H. (ed), Robust Statistics, Data Analysis, and Computer Intensive Methods. Lecture Notes in Statistics, Springer-Verlag 109* (1996), 17–34. Proceedings of the workshop in honor of Peter J. Huber.

[4] Bond, N. W. Impairment of shuttlebox avoidance-learning following repeated alcohol withdrawal episodes in rats. *Pharmacology, Biochemistry and Behavior 11* (1979), 589–591.

[5] Cantoni, E., y Ronchetti, E. Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association 96* (2001), 1022–1030.

[6] Croux, C., y Haesbroeck, G. Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis 44* (2003), 273–295.

[7] Gervini, D., y Yohai, V. J. A class of robust and fully efficient regression estimators. *The Annals of Statistics 30* (2002), 583–616.

[8] Gonçalves, L., Subtil, A., Oliveira, M. R., y Bermudez, P. ROC curve estimation: An overview. *REVSTAT - Statistical Journal 12*, 1 (2014), 1–20.

[9] Greco, L., y Ventura, L. Robust inference for the stress-strength reliability. *Statistical Papers 52* (2011), 773–788.

[10] Inácio de Carvalho, V., Jara, A., Hanson, T. E., y de Carvalho, M. Bayesian non-parametric ROC regression modeling. *Bayesian Analysis 8* (2013), 623–646.

[11] Janes, H., Longton, G., y Pepe, M. S. Accommodating covariates in receiver operating characteristic analysis. *The Stata Journal 9*, 1 (2009), 17–39.

[12] Maronna, R. A., Martin, R. D., Yohai, V. J., y Salibián-Barrera, M. *Robust statistics: theory and methods (with R)*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2019.

[13] McCullagh, P., y Nelder, J. A. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1989.

[14] Murrone, N. *Curvas ROC con presencia de covariables: un estudio de sensibilidad*. Tesis de Licenciatura, 2019. Disponible en: `http://cms.dm.uba.ar/academico/carreras/licenciatura/tesis/2019/`.

[15] Norton, S. J., Gorga, M. P., Widen, J. E., Folsom, R. C., Sininger, Y., Cone-Wesson, B., Vohr, B. R., Mascher, K., y Fletcher, K. Identification of neonatal hearing impairment: evaluation of transient evoked otoacoustic emission, distortion product otoacoustic emission, and auditory brain stem response test performance. *Ear & Hearing 21*, 5 (2000), 508–528.

[16] Pardo-Fernández, J. C., Rodríguez-Álvarez, M. X., y Van Keilegom, I. A review on ROC curves in the presence of covariates. *REVSTAT Statistical Journal 12*, 1 (2014), 21–41.

[17] Pepe, M. S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Stastistical Science Series. Oxford University Press, New York, 2003.

[18] Pregibon, D. Logistic regression diagnostics. *The Annals of Statistics 9* (1981), 705–724.

[19] Rodríguez-Álvarez, M. X., Tahoces, P. G., Cadarso-Suárez, C., y Lado, M. J. Comparative study of ROC regression techniques: Applications for the computer-aided diagnostic system in breast cancer detection. *Computational Statistics and Data Analysis 55* (2011), 888–902.

[20] Seber, G. A. F., y Lee, A. J. *Linear Regression Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, New Jersey, 2003.

[21] Walsh, S. J. Limitations to the robustness of binormal ROC curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power. *Statistics in Medicine 16* (1997), 669–670.

[22] Wedderburn, R. W. M. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss–Newton Method. *Biometrika 61* (1974), 439–447.

[23] Yohai, V. J. High breakdown–point and high efficiency estimates for regression. *The Annals of Statistics 15* (1987), 642–656.

[24] Yohai, V. J., y Maronna, R. A. Asymptotic behavior of M-estimates for the linear model. *The Annals of Statistics 7* (1979), 258–268.